Artificial Intelligence, Politics, and Political Science

Artificial intelligence (AI) represents perhaps the most consequential technological shift for political science in a generation, with the potential to reshape most domains of politics and political inquiry. Produced by the American Political Science Association's (APSA) Presidential Task Force on AI, Politics, and Political Science, this volume offers the first comprehensive, discipline-wide assessment of AI's effects. Drawing on contributions from over sixty leading scholars, it examines how AI is transforming political campaigns, election administration, information ecosystems, public opinion research, geopolitics and national security, policymaking, political theory, economic inequality, racial and gender politics, and political science pedagogy and methods. Each chapter delivers rigorous empirical assessment alongside implications for researchers, practitioners, and policymakers. A central argument runs throughout: AI's political impact is shaped not by technological capability alone but also by the regulatory frameworks, institutional contexts, and political choices governing its deployment. This title is also available as Open Access on Cambridge Core.

Nathaniel Persily is James B. McClatchy Professor of Law at Stanford Law School, where he also has appointments in the Departments of Political Science and Communication and is Co-Director of the Stanford Law AI Initiative. He is a leading expert on election law, democracy, and the regulation of technology.

Joshua A. Tucker is Julius Silver, Roslyn S. Silver, and Enid Silver Winslow Professor of Politics at New York University, where he is also Co-Director of the Center for Social Media, AI, and Politics (CSMAP) and Director of the Jordan Center for the Advanced Study of Russia. His research focuses on the intersection of the digital information environment, including social media and AI, and politics, as well as the comparative politics of mass political behavior.

PROOFS

PROOFS

Tucker_9781047774734_C000.indd 2 04-Sep-26 9:54:51 PM

Artificial Intelligence, Politics, and Political Science

Edited by
NATHANIEL PERSILY
Stanford University

JOSHUA A. TUCKER
New York University

PROOFS

CAMBRIDGE
UNIVERSITY PRESS



Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

Cambridge University Press is part of Cambridge University Press & Assessment,
a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of
education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781047774734
DOI: 10.1017/9781047774635

© Nathaniel Persily and Joshua A. Tucker 2027

This publication is in copyright. Subject to statutory exception and to the provisions of
relevant collective licensing agreements, with the exception of the Creative Commons
version the link for which is provided below, no reproduction of any part may take place
without the written permission of Cambridge University Press & Assessment.

An online version of this work is published at doi.org/10.1017/9781047774635 under
a Creative Commons Open Access license CC-BY-NC-ND 4.0 which permits re-use,
distribution and reproduction in any medium for non-commercial purposes providing
appropriate credit to the original work is given. You may not distribute derivative works
without permission. To view a copy of this license, visit https://creativecommons.org/
licenses/by-nc-nd/4.0

When citing this work, please include a reference to the DOI 10.1017/9781047774635

First published 2027

Cover images:
Robot Arm: wildpixel
Soldier & Drone: Golden Sikorka
Capitol Hill: A-Digit
Map: RomanOkopny
All from Getty Images

A catalogue record for this publication is available from the British Library

Library of Congress Cataloging-in-Publication Data

ISBN 978-1-047-77473-4 Hardback
ISBN 978-1-047-77472-7 Paperback

Cambridge University Press & Assessment has no responsibility for the persistence
or accuracy of URLs for external or third-party internet websites referred to in this
publication and does not guarantee that any content on such websites is, or will
remain, accurate or appropriate.

For EU product safety concerns, contact us at Calle de José Abascal,
56, 1°, 28003 Madrid, Spain, or email eugpsr@cambridge.org.

PROOFS

For Aaron, Noah, Sasha, and Mattie

PROOFS

PROOFS

Tucker_9781047774734_C000.indd 6 04-Sep-26 9:54:51 PM

Contents

List of Figures and Tables *page* ix
List of Contributors xi
Foreword xvii
Taeku Lee
Acknowledgments xxi

Introduction: Artificial Intelligence, Politics, and Political Science 1
Nathaniel Persily and Joshua A. Tucker

1 Artificial Intelligence and Democracy: Campaigns, Elections, Movements, and Deliberation 26
Bailey Flanigan, Florian Foos, Archon Fung, and Charles Stewart III

2 Easy to Produce, Hard to Persuade: The Asymmetric Effects of AI on the Online Information Ecosystem 53
Brendan Nyhan, Jennifer Pan, Alexandra Siegel, and Yamil Velez

3 Public Opinion in the Age of AI 78
Joshua D. Clinton, Soubhik Barari, Ethan Busby, Trent D. Buskirk, Raymond Duch, Anna-Carolina Haensch, D. Sunshine Hillygus, Courtney Kennedy, Kevin Munger, Doug Rivers, and Sean Westwood

4 AI, the Public Sector, and Policymaking 109
Baobao Zhang, Diane Coyle, Jae Yeon Kim, Johannes Himmelreich, and Mila Gascó-Hernandez

5 AI, Race, and Politics 152
Rachel Gillum, Gregory Leslie, and Cara Wong

6 AI, Gender, and Politics 195
Dawn Teele, Shira Pindyck, and Sophia Lipkin

PROOFS

Tucker_9781047774734_C000.indd 7 04-Sep-26 9:54:51 PM

7 AI's Economy and Its Political and Institutional Consequences 209
Carles Boix, Michael Becher, Valentina González-Rostani, and Daniel Stegmueller

8 AI, Geopolitics, and National Security 247
Sarah Kreps, Ben Buchanan, Michael Horowitz, and Erica Lonergan

9 AI and Political Theory 282
Linda Eggert, Ting-an Lin, Jeffrey Howard, Lorenzo Manuali, and Rob Reich

10 AI and Research Methods 313
Christopher Barrie, Lisa P. Argyle, James Bisbee, Michael Heseltine, Christopher Lucas, Jon Mellon, Alexis Palmer, Margaret Roberts, and Arthur Spirling

11 Teaching and Learning: Political Science in the Era of AI 365
John Ishiyama, Christine Cahill, Jennifer De Maio, Stefan E. Kehlenbach, Sing Hui Lee, Steven Michels, Charles C. Turner, and Nicole Wu

Index 397

PROOFS

Figures and Tables

PROOFS

FIGURES

3.1 Selected uses of AI in survey research *page* 84
4.1 Share of federal AI use cases by development mode, among cases with known status 117
4.2 Traditional accountability relationships in public administration 122
7.1 Comparison of three popular AI-exposure measures 214
7.2 Capital expenditures of top five technological companies, 2015–25 225
7.3 AI exposure by income decile in the US 227
7.4 District-level AI exposure and partisan vote in the US 228
10.1 Pseudo-code: traditional agent-based model and generative agent-based model architectures 336

TABLES

4.A Taxonomy of AI use by local, regional/state, and national governments with examples 137
5.1 Mapping the AIMS four-question framework to existing methods standards 178
7.1 Sector-specific exposure measures 215
7.A Overview of task and automation exposure objective measures 235
11.1 Number of articles and papers addressing generative AI per year, in primary APSA outlets on teaching and learning 371
11.2 Sample rubric-assessing digital literacy with AI tools in political science 385
11.A AI and teaching papers from APSA annual meetings (including TLC at APSA) 389

Tucker_9781047774734_C000.indd 9 04-Sep-26 9:54:51 PM

PROOFS

Contributors

Nathaniel Persily
James B. McClatchy Professor of Law
Stanford Law School

Joshua A. Tucker
Julius Silver, Roslyn S. Silver, and Enid Silver Winslow Professor of Politics
New York University

Lisa P. Argyle
Associate Professor of Political Science
Purdue University

Soubhik Barari
Senior Research Methodologist
National Opinion Research Center (NORC) at the University of Chicago

Christopher Barrie
Assistant Professor of Sociology
New York University

Michael Becher
Professor of Political Science
IE University

James Bisbee
Assistant Professor of Political Science
Vanderbilt University

Carles Boix
Robert Garrett Professor of Politics and Public Affairs
Princeton University

PROOFS

Tucker_9781047774734_C000.indd 11 04-Sep-26 9:54:51 PM

Ben Buchanan
Dmitri Alperovitch Assistant Professor
Johns Hopkins University, School of Advanced International Studies

Ethan Busby
Assistant Professor of Political Science
Brigham Young University

Trent D. Buskirk
Professor and Provost Data Science Fellow
Old Dominion University

Christine Cahill
Associate Teaching Instructor and Undergraduate Director, Department of Political Science
Rutgers University

Joshua D. Clinton
Abby and Jon Winkelried Professor of Political Science
Vanderbilt University

Diane Coyle
Bennett Professor of Public Policy
University of Cambridge

Jennifer De Maio
Associate Professor of Political Science
California State University, Northridge

Raymond Duch
Official Fellow, Nuffield College; Director, Nuffield Centre for Experimental Social Sciences (CESS)
University of Oxford

Linda Eggert
Assistant Professor of Philosophy
Stanford University

Bailey Flanigan
Theodore T. Miller Career Development Assistant Professor of Political Science and Electrical Engineering and Computer Science
Massachusetts Institute of Technology

Florian Foos
Associate Professor in Political Behaviour
London School of Economics and Political Science

Archon Fung
Winthrop Laflin McCormack Professor of Citizenship and Self-Government
Harvard Kennedy School

PROOFS

Mila Gascó-Hernandez
Associate Professor of Public Affairs and Policy; Research Director, Center for Technology in Government
University at Albany, State University of New York

Rachel Gillum
Vice President of Ethical and Humane Use of Technology
Salesforce

Valentina González-Rostani
Assistant Professor of Political Science and International Relations
University of Southern California

Anna-Carolina Haensch
Assistant Research Professor, Joint Program in Survey Methodology
University of Maryland and Ludwig Maximilian University of Munich

Michael Heseltine
Postdoctoral Researcher, Department of Sociology
University of Oxford

D. Sunshine Hillygus
Professor of Political Science and Public Policy; Director, Duke Initiative on Survey Methodology
Duke University

Johannes Himmelreich
Associate Professor, Maxwell School of Citizenship and Public Affairs
Syracuse University

Michael Horowitz
Richard Perry Professor of Political Science; Director, Perry World House
University of Pennsylvania

Jeffrey Howard
Professor of Political Philosophy and Public Policy
University College London

John Ishiyama
University Distinguished Research Professor of Political Science
University of North Texas

Stefan E. Kehlenbach
Assistant Professor of Political Science, Rockefeller College of Public Affairs and Policy
University at Albany, State University of New York

Courtney Kennedy
Vice President of Methods and Innovation
Pew Research Center

PROOFS

Jae Yeon Kim
Assistant Professor of Public Policy
University of North Carolina at Chapel Hill

Sarah Kreps
John L. Wetherill Professor of Government; Director, Tech Policy Institute
Cornell University

Sing Hui Lee
Lecturer and Faculty Advisor, Political Science and International Studies Program
University of North Texas

Taeku Lee
Bae Family Professor of Government
Harvard University

Gregory Leslie
Assistant Professor of Political Science
The Ohio State University

Ting-an Lin
Assistant Professor of Philosophy
University of Connecticut

Sophia Lipkin
PhD Student in Political Science
University of Chicago

Erica Lonergan
Assistant Professor, School of International and Public Affairs
Columbia University

Christopher Lucas
Associate Professor of Political Science
Washington University in St. Louis

Lorenzo Manuali
PhD Candidate in Philosophy
University of Michigan

Jon Mellon
Co-Director, British Election Study; Staff Data Scientist, Pinterest

Steven Michels
Professor and Chair, Department of Political Science and Global Affairs
Sacred Heart University

PROOFS

Kevin Munger
Assistant Professor and Chair of Computational Social Science
European University Institute

Brendan Nyhan
James O. Freedman Presidential Professor of Government
Dartmouth College

Alexis Palmer
Assistant Professor of Political Science
Tulane University

Jennifer Pan
Professor of Communication, Senior Fellow at The Freeman Spogli Institute
Stanford University

Shira Pindyck
Visiting Scholar, Perry World House
University of Pennsylvania

Rob Reich
Professor of Political Science; Associate Director, Institute for Human-Centered Artificial Intelligence (HAI)
Stanford University

Doug Rivers
Professor of Political Science, Stanford University; Chief Scientist, YouGov

Margaret Roberts
Professor of Political Science
University of California, San Diego

Alexandra Siegel
Associate Professor of Political Science
University of Colorado Boulder

Arthur Spirling
Class of 1987 Professor of Politics
Princeton University

Daniel Stegmueller
Professor of Political Science
Duke University

Charles Stewart III
Kenan Sahin Distinguished Professor of Political Science
Massachusetts Institute of Technology

PROOFS

Dawn Teele
Stavros Niarchos Foundation (SNF) Agora Professor of Political Science
Johns Hopkins University

Charles C. Turner
Professor, Department of Political Science and Criminal Justice
California State University, Chico

Yamil Velez
Assistant Professor of Political Science
Columbia University

Sean Westwood
Associate Professor of Government; Director, Polarization Research Lab
Dartmouth College

Cara Wong
Professor of Political Science and Asian American Studies
University of Illinois Urbana-Champaign

Nicole Wu
Assistant Professor of Politics and Public Administration
University of Hong Kong

Baobao Zhang
Maxwell Dean Associate Professor of the Politics of AI Maxwell School of Citizenship and Public Affairs
Syracuse University

PROOFS

Foreword

In his 1959 Rede Lecture, "The Two Cultures and the Scientific Revolution," C. P. Snow bemoaned the yawning gap that he saw between scientists, whose work he equated with progress and "the only hope for the poor," and literary intellectuals, who he conventionalized as "natural Luddites." Snow's Rede Lecture introduced "two cultures" to our lexicon and struck a chord not just for its artful invective. In this "gulf of mutual incomprehension" between scientists and intellectuals, between technology and humanity, Snow saw "hostility and dislike, but most of all lack of understanding."

Today, we confront an even more forbidding "two cultures" moment than what motivated Snow a half century ago, one that could redefine governance, global conflict, economic development, and human progress for the foreseeable future. Over the shockingly brief span of a few decades, artificial intelligence (AI) has emerged as the next epochal technological change. The "scientists" of our time champion a brave new frontier of human possibility and "political superintelligence" (see Andy Hall, "Building Political Superintelligence" [March 26, 2026]. Accessed online April 10, 2026 at: https://freesystems.substack.com/p/building-political-superintelligence), while the "intellectuals" forewarn of "cognitive surrender" and existential threats to humanity as we know it, even to our understanding of what it means to be human.

Whether promise or peril, AI is a technological genie that cannot be put back into the proverbial bottle, especially with the recent explosive emergence of generative AI tools. ChatGPT, to take the most obvious current example, reached 100 million users within a month of its launch in November 2022, perhaps the most rapid adoption of any technology in measured history (Dan Milmo, "ChatGPT Reaches 100 Million Users Two Months after Launch," *The Guardian* (February 2, 2023). Accessed online April 10, 2026 at: www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app). By one recent accounting, within a mere two years of

PROOFS

Tucker_9781047774734_C000.indd 17 04-Sep-26 9:54:51 PM

ChatGPT's launch, the quantity of AI-generated content on the internet has surpassed that of human-generated content (https://graphite.io/five-percent/more-articles-are-now-created-by-ai-than-humans). AI now seems poised to alter – possibly radically – virtually every aspect of society, from the economy to culture, education, information, and politics.

Academia is not spared from such change and neither is the discipline of political science. AI's impact spans the gamut from forcing a dramatic rethinking of our pedagogic practices and our metrics of evaluation in the classroom all the way to reckoning with a possible future of "academic slop" where AI could enable individual researchers to generate thousands of passable empirical papers each year (on slop, see Seva Gunitsky's "The Age of Academic Slop Is Upon Us" [January 13, 2026]. Accessed online April 10, 2026 at: https://hegemon.substack.com/p/the-age-of-academic-slop-is-upon. See also Graham Straus and Andy Hall, "How Accurately Did Claude Code Replicate and Extend a Published Political Science Paper?" [January 9, 2026]. Accessed online at: https://andrewbenjaminhall.com/pdfs/Straus_Hall_Claude_Audit.pdf). For political science, there is the additional aspect of AI's potential to influence the very phenomena we study or to add emergent phenomena to our research agendas, such as new frontiers of algorithmic governance, AI-enhanced propaganda swarms, AI-assisted cyberterrorism, all the way to the possibility of private firms like Amazon, Alphabet, and Microsoft wielding sovereign power over nation-states.

This volume is the result of the recognition that there is a singular urgency for political science to take stock of AI's potential impact on the discipline and on the profession. As the President of the American Political Science Association (APSA), I launched an APSA Task Force on AI and Political Science in 2025 to mobilize an all-hands-on-deck effort. I persuaded two sterling political scientists – Nathaniel Persily and Joshua Tucker – to answer the call of duty to the discipline and serve as co-chairs of the Task Force. We then assembled an extraordinary group of scholars to join us. The Task Force, and its contributions in this volume, represent an interdisciplinary initiative, bridging all fields of political science – empirical and philosophical – and fields beyond political science to examine the impact of AI on politics and our profession.

At the time that I launched the Task Force, my untested hypothesis was that political scientists had sorted into three "cultures" on AI. The first being a culture of embrace among those early and eager adopters who took to AI models and tools like a five-year old to their first Lego set. The second being a culture of denial among the Never AI-ers, who would just as eagerly teletransport to an alternate universe where AI did not exist and where its future discovery would be unimaginable. And the third, probably biggest, being a bloated middle culture of those who are infelicitously torn on alternate days between being AI-curious and AI-skeptical, who have dipped their toes into its potential but are not ready to fully reckon with the scope of change that AI portends.

PROOFS

In his essay, C. P. Snow entreated a closing of the gap between cultures so that the pressing problem of his time – global inequality – might be tackled together. In the interceding decades since his essay, global inequality has only deepened. To that wicked problem, we now add newer menacing threats from polarized polities and backsliding democracies to the fragmentation of shared facts and the existential threats from climate change, all in a changing political environment of diminishing public investments in knowledge and research. AI may ameliorate these threats, or it may aggravate and accelerate them. What is clear in any scenario is that AI is upon us and that AI will impact our ability to teach, research, and learn about any of these pressing problems.

My aspirations for the Task Force were thus to agenda-set for a reckoning between cultures in political science on AI. Concretely, I hoped to extract from the collective wisdom of Task Force members an empirical foundation and a set of apposite tools and theories to meet AI's challenges and opportunities. This volume far exceeds my highest expectations. Readers will of course judge for themselves. What should nonetheless be clear by any generous reading is that Persily and Tucker have masterfully orchestrated a remarkably broad and insightful set of contributions that will inform all of us on how AI may alter how we study politics, how we teach about politics, and what politics itself will look like.

Taeku Lee

PROOFS

PROOFS

Tucker_9781047774734_C000.indd 20 04-Sep-26 9:54:51 PM

Acknowledgments

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 American Political Science Association (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 University Press, including John Haslam, Claire Sissen and Carrie Parkinson, and from Lumina Datamatics, Srishti Prakash and Swati Kumari, who shepherded this volume to publication. We would also like to thank the Stanford Law AI Initiative and New York University's (NYU) Center for Social Media, AI, and Politics for providing institutional support. Funding to make this volume available Open Access was provided by Stanford Law School and the American Political Science Association.

PROOFS

Tucker_9781047774734_C000.indd 21 04-Sep-26 9:54:51 PM

PROOFS

Introduction

Artificial Intelligence, Politics, and Political Science

Nathaniel Persily and Joshua A. Tucker

PROOFS

1.1 INTRODUCTION

Artificial intelligence (AI) is reshaping politics and political science, just as it is transforming other social phenomena and their associated academic fields. As with the larger policy debates over AI and its social impact, attitudes toward the newest version of this technology range from utopian to dystopian, with many also alleging the technology is overhyped, at least in the short term. With the pace of technological and political change nearly outpacing the capacity of (human) academics to analyze these trends, any endeavor to take stock of where things stand for AI and politics in the summer of 2026 is necessarily fraught. Although the trajectory remains uncertain, a volume like the one that follows provides a critical snapshot of the state of the field as it begins to grapple with the multifaceted questions of AI's relationship to politics and research.

It was against this background, in the fall of 2025, that Taeku Lee, then President of the American Political Science Association (APSA), commissioned a Presidential Task Force on AI, Politics, and Political Science, naming us as co-chairs. We then identified the topics that would comprise the report and identified potential committee chairs for each topic. Upon agreeing to serve on the task force, each committee chair then assembled committees to write the chapters in this volume, which collectively represent the "Report" of the Task Force. Admittedly, as an edited volume in which different authors express varying perspectives on the themes related to AI and politics, this Report differs from other task force reports. However, with (humble) recognition of the variety of views on the relevant questions at this early stage of technological

AI Disclosure Statement for this chapter: AI was not used in the construction of this introduction, except to provide chapter summaries for the "Overview of Chapters" section, which the authors subsequently rewrote and we edited.

Tucker_9781047774734_Introduction.indd 1 07-Sep-26 9:34:26 AM

development, we thought that surfacing these disagreements was itself a form of service to the discipline. Moreover, the topic is so vast, multifaceted, and quickly changing that crafting a volume touching on many, if admittedly not all, the subfields of political science in the space of six months required that we grant considerable autonomy to each committee. In that spirit, members of the Task Force speak only for themselves in the chapters that they have authored, and the volume represents a collection of views rather than a definitive assessment of the appropriate place of AI in the study and practice of politics.

Our goal with this volume, therefore, is not to offer predictions about the future of AI but, rather, to map the state of existing research, identify the most pressing questions, and establish a foundation for the rigorous empirical study of AI's effects on politics and governance. None of this is to suggest that we must resign ourselves to mere speculation in assessing the impact of AI: AI is already affecting the practice and study of politics in identifiable ways. Regulatory bodies throughout the United States and around the world have already begun to grapple with the use of generative AI in political campaigning, for instance. Administrative agencies are rolling out AI tools to deal with case backlogs and to improve public services. The issue of AI in warfare is becoming increasingly relevant in armed conflicts in Ukraine, the Middle East, and around the world. And political scientists, like other scientists and social scientists, are employing new AI tools to study a range of political phenomena – so much so, in fact, that we are already seeing glimpses of a transformed profession in which researcher productivity outstrips the capacity of traditional institutions (such as peer-reviewed journals) to curate and to present political science to a larger audience.[1]

1.2 THE SCOPE OF THIS VOLUME

One of the challenges with a volume like this is that the topic is potentially so capacious that disagreements inevitably emerge as to what phenomena are in and out of scope. Different people mean different things when they reference AI, and the range of topics within political science that might be relevant to AI arguably covers the entire discipline. We hope that this volume begins a conversation about AI and the discipline of political science and that omissions are filled in by later work.

We begin with the question as to what we mean by AI. At its most general and abstract, the concept could include all forms of machine learning and digital algorithms, but surely that definition is both overbroad and underinclusive. In current popular imagination, largely shaped by consumer experiences with ChatGPT, generative AI (in the form of chatbots and image generators) has been the primary focus of any discussion concerning AI and politics, especially as it pertains to issues related to deep fakes and the generation of AI content

[1] See Munger et al. (2026).

PROOFS

on social media. However, generative AI is just one incarnation of AI (let alone politically relevant AI), the definition of which must also include the brave new world of agentic systems and "embodied AI" in the form of robotics, autonomous vehicles, and drones. In the national security context, for example, these latter types of AI may be most significant.

The definition of AI eventually adopted by the European Union in its AI Act is a typical one: "a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments."[2] However, the EU arrived at that definition only after several years of debate and after having gone back to the drawing board following the explosion of generative systems following the launch of ChatGPT. This definition is as good as any, so long as we keep in mind that the world of politically relevant AI includes all of the related systems (robotics, etc.) described earlier. For purposes of this volume, we do not feel the need to settle any debate over the proper definition of AI, but, rather, we want to emphasize that it includes a family of technologies that are ever-evolving and that focusing only on the most popular AI products today (such as chatbots) ignores a diverse array of tools with political relevance.

As difficult as it may be to define AI, it is equally challenging to fence off the terrain of politics and political science that might be relevant to, or affected by, AI. In a sense, any of AI's large-scale social benefits or harms could properly be within the scope of "politics." To the extent AI poses existential risks – such as through increased probabilities of the development of rogue chemical, biological, radiological, or nuclear (so-called CBRN risk) weapons or through acceleration of climate change, massive labor displacement, or a Skynet-Terminator scenario of robot domination of humans – all those developments are properly described as "political." Similarly, if AI helps cure diseases like cancer, dramatically increases life expectancy and the quality of life, or leads to massive economic growth enabling millions to lead a life of leisure, politics will be affected.

These potential impacts go far beyond the heuristics that many bring to thinking about AI and politics. The mental model many analysts bring to these questions was developed in the social media age, when technology-related concerns about misinformation, polarization, privacy violations, and political manipulation were top of mind. This volume deals with those issues, but, as argued later in this introduction, we worry that an excessive focus on AI's relationship to social media understates the relevant impact of AI on politics. In contrast, we have attempted to include in this volume representatives from as many different subfields of political science as is reasonably possible. To be

[2] See EU AI Act, Article 3 (1) (European Union, 2024).

PROOFS

sure, not every subfield of political science is represented, but we hope that this book provides a springboard for similar volumes, perhaps each focusing on a particular subdiscipline of political science.

Finally, this volume has the dual ambition of not merely describing the politics of AI but also reflecting on what AI means for the practice of political science. Here, too, we do not cover all possible AI-related impacts on the job of a political scientist. Indeed, depending on how capable AI systems become, we can only begin to speculate what these technological developments might mean for the future of universities or the role of humans as researchers and teachers (let alone the expectations of students as to what is worth learning in the AI age). The chapters that focus on the profession consider the impact of AI on traditional forms of political science research, writing, publishing, and teaching: the methods and tools we use to perform that job as it has evolved over the last century or so. We recognize, however, that the tectonic plates underlying university research and teaching are shifting, and that in a few decades, we might look back on this volume with a nostalgia for the traditional job of a political scientist, which AI changed into a fundamentally different occupation.

But the stakes here are not just professional. To the extent that political science provides a framework for societies to address distributions of power with shared evidence and methods, it is incumbent upon political science as a discipline to wrestle with the way AI will change our *standards* of methods and evidence. If AI fundamentally changes the way we produce knowledge as a field – which is still a big if – then the quality of the arguments we can present to the public may change as well. Understanding the nature of those changes is another goal of this volume.

1.3 TEMPORAL, GEOGRAPHIC, AND DATA CHALLENGES IN AI RESEARCH

We would be remiss not to acknowledge the significant challenges that attend any effort to assess the influence of a fast-developing technology with global implications on a broad array of political phenomena and the academic study of them. Even as scholars who frequently write about technology and politics, we find it particularly challenging to keep up with the latest AI developments, let alone to publish analyses of their political implications in a timely manner.

Between the time our contributors drafted their chapters and the time this volume reaches readers, new models have been released (and, in the case of Anthropic's Fable, unreleased and then re-released), new capabilities have emerged, and some of the specific examples cited here may already feel dated. The lag between technological development and scientific analysis of its impact is an unavoidable feature of the landscape. For example, after the authors submitted their chapters, we have witnessed dramatic developments in the deployment of agentic AI systems (such as Claude Code and Codex) with implications for political science research, let alone the larger economy. We have also seen

PROOFS

the emergence of new political controversies, such as those involving Anthropic and the U.S. Government, that seem to present unique and fundamental questions related to the "politics of AI." Several of our contributors note the difficulty of writing about a technology with capabilities that shift on the order of weeks and months rather than years. We have, therefore, tried to focus the volume on analytical frameworks, conceptual distinctions, and empirical findings that will retain their value even as the specific technologies evolve.

A second challenge is the geography of AI development and concomitant research on AI. The leaders in the race to develop frontier models are overwhelmingly located in two countries: the United States and China. Relatedly, English and Chinese tend to dominate the training data used in the production of frontier models. Moreover, a great deal – although not all – of the literature that is reviewed in this volume comes from analyses of AI in the context of the United States, be it in terms of policy, the information environment, security questions, or other areas. This is perhaps not surprising at this point in the field's development, but also both an important blind spot to be recognized and opportunity for future research. Models may perform differently in low resource languages – and with predictable consequences.

Thus, to the extent that this volume focuses on the existing research on AI and politics, which has a pronounced Western bias, it necessarily leaves out much of the critical work done on these topics in the rest of the world. This volume, incomplete though it may be, should also be read as a clarion call for a geographic broadening of the relevant research agenda. To the extent that AI "lowers the cost" of politics – that is, by reducing the costs of carrying out fundamental political activities, from running political campaigns to surveilling political opponents – we might expect its impact to be felt even more profoundly in developing countries. The field of political science needs to account for the global impact of this technology and dedicate scholarly resources to ensure an inclusive assessment of the far-reaching political impacts of AI.

A third challenge, and one that will be familiar to readers of our earlier work on social media (Persily and Tucker, 2020), concerns the problem of data access. The training data, model architectures, and user interaction logs that would be most valuable for studying AI's political effects are almost entirely locked behind corporate firewalls. We are, once again, in a position where a small number of extraordinarily powerful private firms control access to the data necessary to assess the impact of a new technology on society. This is not to say we do not have tools available to study AI; as we explain in more detail later in this chapter, we may actually be in a better place, due to the business models of AI firms, to observe the outputs of these models than we were for, say, social media algorithms. But we remain hamstrung in our ability to observe how human beings interact with these AI systems – to say nothing of the inner workings of the systems themselves.

PROOFS

1.4 LESSONS FROM THE STUDY OF SOCIAL MEDIA

When analyzing the likely effect of AI on politics, commentators naturally gravitate to the most recent example where technological innovation has been blamed for political change: social media. The reliance on this paradigm is unsurprising, as some of the most ardent criticisms of AI focus on its nexus to social media, through generation and propagation of deep fakes or hate speech, or other familiar dangers related to privacy, children, addiction, or emotional well-being. However, AI is a general purpose technology with implications for the economy and society that far exceed even the direst warnings (or highest praise) about social media. For all of the heated criticisms, no one ever feared that social media would lead to mass unemployment or human extinction. Although we agree that the rise of social media can be a useful heuristic for thinking about AI, it is important that we do not simply default to lessons learned about social media and assume things will be the same with AI. Wholesale transference of the political anxieties developed with respect to social media misrepresents the potential, for both good and ill, of this newest family of technologies to have a substantial impact on politics and governance.

The levels of skepticism and concern Americans express relating to AI are, no doubt, a product of the unrelenting criticism (both warranted and unwarranted) of social media over the past decade. It should be noted that in the early days of social media, the technology was hailed as a "liberation technology" and lauded for its likely democratizing impacts (Diamond, 2010). However, over the past decade, Americans have come to distrust both the government and Big Tech. Put those two together and it is no surprise that large majorities in the United States express concern about AI and feel that it will do more harm than good.[3] Those anxieties have now led to explicit acts of political violence, as anti-AI activists threw a molotov cocktail at the house of OpenAI CEO Sam Altman and fired gun shots at the home of an Indianapolis councilman with a note placed on his doorstep saying "no data centers." If the frame that people bring to these questions is one forged from worries, for example, that social media caused untold political harm around the globe, we should not be surprised by the level of political anxiety accompanying the rollout of AI. Nor should we be surprised that most other countries with populations expressing more favorable opinions of social media also have more favorable views about AI.[4]

Next we consider four characteristics of social media research in political science to briefly assess the extent to which we think these prior experiences with social media are or are not likely to inform our understanding of AI and its relationship to the study of politics. These are the ways in which the new technology (formerly social media, now AI) (1) plugs into existing research

[3] Quinnipiac University (2026).

[4] See Wike et al. (2022); Capstick (2025).

PROOFS

topics in political science; (2) creates new methodological challenges; (3) drives methodological innovation; and (4) creates new challenges related to data access.

I.4.1 Plugging into Existing Political Science Research Topics

In addition to functioning as a new source of data, social media increasingly attracted attention from political scientists because of its role as a potentially important independent variable affecting important outcomes of interest in political science. Social media originally burst into the collective consciousness of scholars studying protest, first in authoritarian regimes (e.g., Arab Spring) and then later in democracies (e.g., Occupy Wall Street). However, it was not long before the number of political topics where the study of social media seemed relevant grew substantially, coming over time to include voting behavior, public opinion formation, elite communication and campaigning, foreign influence campaigns, and, perhaps most significantly, political polarization and the political information environment. As this volume demonstrates, AI seems poised to follow a similar trajectory to social media in this regard, as we suspect, political scientists will increasingly find themselves wrestling with substantive questions related to AI. Although the direct effect of AI on the political topics of interest may not be quite as clear as was the case for social media, the wide range of topics covered in this volume illustrates that AI may prove in the long term to be even more consequential in terms of the sheer number of political arenas (e.g., defense, courts, and procurement, in addition to all of the more obvious information environment implications) in which its impact may be felt.

I.4.2 Methodological Challenges

Social media data did not look like traditional political science data in so far as they did not fit neatly into a spreadsheet where all of the cells were filled with numbers. Instead, social media contained text (and eventually images and video) in a networked format, supplemented by multiple levels of metadata about posts, users, content, devices, and so on. In many ways, the rise of AI and, in particular, generative AI simply continues these trends, with text, images, and videos increasingly serving as the units of analysis. Another methodological challenge to studying social media was that the subject of study itself was changing more rapidly than other subjects of political science research. Kevin Munger popularized the term "temporal validity" to address the question of whether the conclusions from a study of, for example, Facebook in 2018 would still be considered valid for understanding Facebook's impact in 2020.[5] The study of

[5] Munger (2023).

PROOFS

AI faces a similar, and potentially even more extreme, challenge in this regard, as the underlying models that are driving the outputs from AI systems seem to be changing at an even faster rate than social media platforms and algorithms. Moreover, the differences in technological capabilities from one frontier model to the next feel much more significant, especially recently, than the changes from one style of algorithmic feed to the next on a social media platform.

1.4.3 Methodological Innovation

The advent of social media went hand in hand with the text-as-data revolution in the field, as high performance computing ushered in an era when rigorous quantitative analysis was now possible using text in addition to numbers. Although not limited to social media data alone, the challenges of working with these data led to political scientists developing skills in machine learning and network analysis. Here perhaps, the difference between social media and AI is most stark. Whereas social media provided new forms of data that spurred methodological innovation, AI itself is a form of methodological innovation. Although drawing the line between machine learning and what is now generically called AI is an imprecise exercise – and, as noted earlier, one in which we have deliberately chosen not to engage in this volume – it is clear that AI, and especially the more recently emerging agentic AI tools, such as Claude Code or Open AI's Codex, has the potential to radically change many aspects of the research process in which political scientists engage. These changes may be orders of magnitude more dramatic – and may happen much more rapidly – than in the past.[6]

1.4.4 The Need for Data Access

Finally, political scientists' attempts to dislodge data from social media companies to study politics have laid the groundwork for contemporary challenges with the AI companies (several of which are, of course, the same companies). As we noted in our previous edited volume (Persily and Tucker, 2020), the reliance on these powerful private companies for some of the most important information related to political phenomena created significant challenges for researchers. These companies are monopolies that control the information relating to basic facts about contemporary society. We wrote at the time:

> As dangerous as these information monopolies may be for purposes of economic competition, such dangers are compounded when only those who work for the firms and share in their corporate missions are able to gain social insights from the data they possess.... [T]he platforms control the information that most richly describes politics and society and therefore the data necessary to make sound judgments across virtually all major policy domains. (Persily and Tucker, 2020, p. 320)

[6] Hall (2026a).

PROOFS

As with social media, the relevant question is not whether the important data will be gathered, but whether the only people who will have access to the data will be employees tied to the profit maximizing mission of the firms controlling the frontier models.

Consequently, the three paths that researchers pursued with respect to social media remain the ones available today: work without the cooperation of the platforms to collect data; work with the cooperation of the platforms to collect data; or work with governments to compel the release of privacy protected data in secure environments or aggregated data released publicly. All three of these approaches have their shortcomings and raise concerns as to whether they can ultimately be successful.[7] Our advice, therefore, remains the same: Researchers need to pursue all of these pathways simultaneously or else we, as a mass public, will risk being seriously under-informed about the possibly massive impact of AI on politics and society.

We do, however, want to highlight one important distinction between data access challenges for social media data and AI data. Yes, both are controlled by the whims of enormously powerful companies. But with social media, it has been possible, through a variety of different tactics, to observe how individuals are interacting with social media platforms: what they are posting, what posts are generating higher levels of engagement, and even in many cases which networks users are choosing to join. Auditing algorithms, around social media feeds, for example, has been much more challenging. Generative AI – and especially for the "closed weight" commercial models – has a similar level of opacity around the inner workings of the technology.

However, generative AI companies are in the business of selling users access to the models, so researchers are better positioned to audit the models through repeated interactions with them. There are, of course, usage limitations and pricing with which researchers need to contend, but, for now at least, there are very real opportunities for researchers to probe how these models respond to different queries. Conversely, though, it is much harder for researchers to observe how *people* are interacting with these new AI platforms, as compared to social media, as most interactions with generative AI (at least for now) are not in public fora but rather in private one-on-one interactions between users and the models. Thus, researchers will need to lean much more heavily into voluntary data donations from users, with all of the associated challenges for inferences, or they will need to rely on the platforms themselves to share user logs with those on the outside, with all of the associated concerns that entails.[8]

[7] For example, the EU Digital Services Act provides a pathway for researchers to demand access to social media data, but European researchers have only begun to take advantage of it and are facing bureaucratic hurdles. No similar program for data access for AI companies currently exists.

[8] As early as 2023, one of us (Tucker) was publicly calling for AI companies to facilitate data donation for research in their chatbot products, but so far there has been no movement in this direction. See Sanderson and Tucker (2023).

PROOFS

1.5 OVERVIEW OF THE CHAPTERS

The volume begins with the substantive domains in which AI is reshaping political life, before concluding with two chapters that examine how AI is transforming the practice of political science itself – its methods and its pedagogy.[9]

The volume opens with Bailey Flanigan, Florian Foos, Archon Fung, and Charles Stewart III's examination of "Artificial Intelligence and Democracy: Campaigns, Elections, Movements, and Deliberation." The authors assess how AI alters what different political actors can do across four critical domains. In campaigns, they distinguish between efficiency gains in back-office operations and the more transformative capacity for "listening at scale." In election administration, they note the potential for AI to relieve resource-constrained officials while cautioning against the risks of vendor dependence and hallucinated voting information. For social movements, the picture is mixed: AI lowers the barrier to producing high-quality content but also empowers activists to analyze government data and coordinate action. Most provocatively, the chapter argues that AI may solve the historic problem of scale in citizen deliberation. It points to experiments like Taiwan's "vTaiwan" initiative and Google DeepMind's "Habermas Machine" as evidence that AI-facilitated deliberation could move from niche experiment to central democratic practice.

In their chapter "Easy to Produce, Hard to Persuade: The Asymmetric Effects of AI on the Online Information Ecosystem," Brendan Nyhan, Jennifer Pan, Alexandra Siegel, and Yamil Velez provide a critical corrective to the prevailing panic about the persuasiveness of AI-generated content. Their central argument is one of asymmetric effects: While AI drastically lowers the cost of producing content, it does not solve the problem of persuading audiences. Most people consume relatively little political news and have strong partisan filters. The direct electoral impact of deepfakes and AI-generated text may, therefore, be less catastrophic than commonly assumed. The chapter instead identifies subtler dangers: the relaxation of constraints for low-capacity actors such as smaller states and extremist groups, the risk that state control over AI development embeds preferred narratives into the technology itself, the degradation of information environments in low-resource languages, and the "liar's dividend," whereby the mere existence of AI-generated content provides a ready-made excuse for dismissing authentic evidence. The chapter concludes by discussing AI's implications for research, both as a tool and as an object of inquiry in its own right.

What does it mean to study public opinion if the responses being studied reflect neither the "public" nor "opinion"? Joshua D. Clinton, Soubhik Barari, Ethan Busby, Trent D. Buskirk, Raymond Duch, Anna-Carolina Haensch,

[9] In an attempt to model the transparency around AI usage for which we are calling later in this chapter, we note that these chapter summaries were originally drafted by Claude and then rewritten, in whole or in part, by the chapter authors.

PROOFS

D. Sunshine Hillygus, Courtney Kennedy, Kevin Munger, Doug Rivers, and Sean Westwood tackle these foundational challenges in "Public Opinion in the Age of AI." The chapter explains how the largest challenges AI poses for the study of public opinion are perhaps not technical, but conceptual. Although AI certainly portends the possibility of intriguing and important refinements – if not wholesale transformations – in how political scientists may choose to measure and characterize public opinion, the chapter also highlights the epistemological challenges that the use of some tools pose. Arguably the deepest risk is not technical errors, but the implication of substituting AI-generated responses for contemporaneous human voices. Given the foundational role that public opinion has for core notions of representation, accountability, and consent, it is profoundly important to raise questions about how the use of AI may affect the very notion of public opinion, in addition to its measurement. Transparency in the use of AI is of utmost importance when its use risks distorting our understanding of public opinion.

Baobao Zhang, Diane Coyle, Jae Yeon Kim, Johannes Himmelreich, and Mila Gascó-Hernandez shift the focus to the state itself, in "AI, the Public Sector, and Policymaking." The chapter presents a taxonomy of governmental AI uses, mapping thirteen categories of activity onto three core governance functions (policymaking, public service delivery, and internal management) with examples from governments worldwide. The authors identify three interconnected challenges: the tension between public service values and algorithmic optimization; the growing information and power asymmetry between governments and private vendors rooted in decades of outsourcing that have hollowed out the state's technical capacity; and the resulting accountability gaps across vertical, horizontal, and diagonal dimensions of democratic governance. Drawing on cases ranging from Michigan's MiDAS unemployment system to Indiana's IBM welfare automation scandal and many others worldwide, they demonstrate what goes wrong when governments deploy AI without clear governance frameworks and strong state capacity. They argue that transparency, accountability, and enforceable constraints are central to a use of AI in the public sector that benefits citizens while reducing risks.

Rachel Gillum, Gregory Leslie, and Cara Wong examine "AI, Race, and Politics," situating artificial intelligence as both a reshaping force in political life and a system that reflects existing structures of racial and ethnic inequality. The chapter argues that bias in AI is often structural rather than incidental, emerging from models trained on data generated in stratified environments, even as AI can, under the right governance and design conditions, expand access and lower barriers to participation. Across domains of state governance, political communication, and electoral administration, the authors trace how AI systems redistribute visibility, voice, and administrative burden in ways that may reinforce or reconfigure racialized patterns of power. The chapter also demonstrates that AI is not only an object of political analysis but an increasingly central component of how political knowledge is produced,

PROOFS

functioning as a measurement instrument whose errors are systematic and often correlated with the groups under study. In response, the authors introduce the AI Measurement Statement (AIMS), a four-question disclosure framework that makes visible how AI systems operate as measurement tools, where they introduce group-differentiated error, and how those risks shape inference. AI Measurement Statement establishes a baseline expectation for how political scientists document and evaluate AI-mediated measurement and defines a standard for transparent, credible, and cumulative research across the discipline.

Dawn Teele, Shira Pindyck, and Sophia Lipkin explore "AI, Gender, and Politics," framing AI as a general-purpose technology that is reshaping the gendered distribution of power. The authors articulate a three-part framework: AI as a tool for data creation, a method of data analysis, and a category of analysis in its own right. As a tool, AI offers powerful new ways to reveal gender inequality – from computer vision analysis of non-verbal communication in debates to natural language processing that detects subtle biases in legislative text. As a category of analysis, the chapter asks how AI will reshape the material conditions of women's lives, raising the prospect of a new "breadwinner backlash" as automation transforms white-collar industries. The chapter also documents the rise of deepfake-enabled gender-based violence as a distinctly new form of political harm.

Carles Boix, Michael Becher, Valentina González-Rostani, and Daniel Stegmueller examine the political and institutional consequences of AI's economic impact in "AI's Economy and Its Political and Institutional Consequences." Unlike previous waves of information technology that primarily displaced routine manual labor, the authors argue, AI threatens to substitute for high-skill, non-routine cognitive tasks – the work of the educated professional class. The chapter presents evidence that exposure measures for AI-related job displacement are surprisingly poorly correlated, yielding contradictory forecasts about which sectors are most vulnerable. The chapter then points out that AI's consequences for employment, wages, and collective bargaining are likely to be mediated by labor-market institutions and the pace of technological adoption. More consequential, the authors argue, is the concentration of capital. The training of frontier AI models requires immense computational resources, accelerating the concentration of economic power in a small number of firms, increasing business investment in fixed assets, and intensifying the incentives of big technology firms in regulatory capture. This fusion of capital concentration and potential labor displacement, they warn, may pose a substantial challenge to the social consensus on which democratic capitalism depends. It may also deepen global inequalities by encouraging reshoring and widening the technological gap between advanced and developing economies.

Sarah Kreps, Ben Buchanan, Michael Horowitz, and Erica Lonergan examine the international dimension of AI in "AI, Geopolitics, and National Security." They frame AI as a system-altering, general-purpose technology

PROOFS

that differs from earlier military innovations in both its diffusion and its integration across civilian and military domains. Unlike nuclear weapons, AI is dual-use, commercially embedded, and widely accessible, making it difficult to separate economic activity from military capability. The chapter argues that competition over AI increasingly turns on control over material and organizational foundations such as semiconductors and compute infrastructure rather than algorithmic breakthroughs alone. Yet rapid diffusion and the difficulty of measuring capabilities generate persistent uncertainty about relative advantage, intensifying security dilemma dynamics. In turn, because AI systems are opaque and resist direct verification, governance is likely to rely on indirect mechanisms such as export controls, standards, and coordination, rather than the formal arms-control regimes of the Cold War.

In "AI and Political Theory," Linda Eggert, Ting-an Lin, Jeffrey Howard, Lorenzo Manuali, and Rob Reich engage with the normative foundations of the discipline. The chapter argues that AI is forcing a convergence of theoretical traditions that have historically operated in isolation. Understanding algorithmic injustice, for example, requires blending insights from distributive justice, relational equality, and epistemic justice. The authors grapple with the "responsibility gap" that emerges when autonomous systems cause harm – can liability be traced to the programmer or does the autonomy of the system create a moral void? And they explore whether AI-facilitated deliberation can enhance democratic legitimacy or whether the mediation of the public sphere by non-human agents inevitably erodes the communicative action that democratic theory requires.

The volume's final two chapters turn to the discipline itself. In "AI and Research Methods," Christopher Barrie, Lisa P. Argyle, James Bisbee, Michael Heseltine, Christopher Lucas, Jon Mellon, Alexis Palmer, Margaret Roberts, and Arthur Spirling provide a comprehensive survey of how large language models (LLMs) are being integrated into the political scientist toolkit. They identify six major applications – annotation and measurement, experimental treatment generation, silicon sampling, generative agent-based modeling, tool-augmented data collection, and the study of LLMs as political objects in their own right – and assess the methodological challenges each presents. The chapter argues that LLMs do not simply make existing methods faster; they alter the standards by which validity, reproducibility, and transparency must be evaluated, and they recommend new practices of documentation and disclosure.

John Ishiyama, Christine Cahill, Jennifer De Maio, Stefan E. Kehlenbach, Sing Hui Lee, Steven Michels, Charles C. Turner, and Nicole Wu close the volume with a discussion of disruption of political science pedagogy in "Teaching and Learning Political Science in the Era of AI." Drawing on surveys of the APSA membership and a systematic review of the emerging literature, the authors document a discipline in transition: Most faculty view AI tools with skepticism but recognize the futility of prohibition. Instead, political science

PROOFS

instructors should focus on teaching students the ethical and responsible use of AI. The chapter argues for a shift from product-oriented to process-oriented pedagogy, moving away from the "disposable assignment" toward scaffolded, project-based learning that makes the process of intellectual creation visible. Crucially, the authors frame AI literacy not merely as a technical skill but as an ethical and civic imperative, one that requires students to grapple with the environmental costs, labor practices, and power asymmetries embedded in these systems.

I.6 CROSS-CUTTING THEMES

Although each of the chapters stands on its own and investigates very different dimensions of the relationship between AI, politics, and political science, we want to end this introduction by identifying some themes that cut across these different topics. The chapters identify several common dynamics relating to AI's effect on politics, as well as common challenges and opportunities that AI presents for political science. We see four themes, each with a dimension relevant to politics and research: (1) the political dynamics posed by the agglomeration of unprecedented power among a select few American corporations, (2) the opacity of AI models and difficulty replicating AI results, (3) inherent biases of AI models and systems, and (4) the need for appropriate benchmarks and standards.

I.6.1 The Unique Power of AI Corporations

The political challenge posed by powerful private actors in capitalist systems does not represent an unfamiliar phenomenon for political scientists. Indeed, depending on how broadly (or historically) one considers these issues, the study of corporate influence on politics stretches back to Karl Marx and Adam Smith, if not further. The impact of corporate power on politics is a recurring issue in research on campaign finance, lobbying, and the administrative state, let alone international relations scholarship on the role multinational corporations play in diplomacy, war, interstate inequality and development, and geopolitics generally.

All that said, the emerging dynamics posed by the power and influence of AI corporations represent an extreme difference in degree, if not in kind. To be sure, we have seen similar concerns raised regarding the immediately preceding generation of American tech companies, such as Meta, Google, Amazon, Apple, Twitter, and Microsoft, as they dominated various parts of the international technology stack and projected their power and values on topics as central to politics as permissible speech. The new dynamics, represented not only by many of these same companies, but also including OpenAI, Anthropic, Nvidia, and others in the US and China, present the specter of companies with technology capable of influencing entire economies and political systems.

PROOFS

Indeed, "influence" may understate the relationship, as these companies will be in the position of providing (or not) everything from the infrastructure to the technological capabilities for all governmental and economic sectors.

The dependence of governments on these corporations seems qualitatively different from classic examples (think Monsanto, Exxon, Dole, or Nestlé) in which critics raised concerns about foreign corporations holding less powerful governments hostage. As Open AI and Anthropic's warnings about the cyberoffensive capabilities of its most recent models attest, these companies are developing technology with such wide- ranging capabilities that they could literally bring down economies and political systems. Consequently, the relationship between the large AI labs and governments has been complex, to say the least. At various points throughout 2026, for example, Anthropic refused to give the federal government access to its models unless it promised not to use them for mass surveillance or autonomous weapons. The company then released its powerful Mythos model to a limited set of trusted partners, as opposed to users generally, given the cybersecurity risks it posed. Later, after Anthropic released a more limited version of Mythos to the public called Fable, the federal government banned its use by non-US nationals, leading Anthropic to withdraw the model for all users since it could not guarantee that only Americans might access the model if released domestically. Then, the government reversed course a few weeks later after Anthropic implemented some security measures, while the Commerce Department still "reserve[ed] the right to reevaluate the decisions made … should circumstances change."[10]

Add to these unique tensions between the companies and governments the prospect of major economic shocks through transformations of the future of work and the labor market, and the politics of AI look quite different from the classic fears of corporate control of politics. With the United States federal government largely opting for a laissez faire approach to regulating these companies – going so far as to threaten other countries that might adopt more aggressive policies while also unilaterally holding up the deployment of models deemed (for undisclosed reasons) to be too dangerous – governments are now placed in a uniquely difficult position of deciding whether to accede to the wishes of these US and Chinese companies or to be left behind in the AI economy with severe domestic political consequences.

The power of these companies is also relevant for political science research. As we noted earlier, these companies control access to data that will be relevant to contemporary politics and policy debates. Researchers are therefore dependent on the frontier labs' decisions regarding what data they choose to make publicly accessible and available to external researchers.

[10] Letter from Howard W. Lutnick, Secretary of Commerce, to Tom Brown, Chief Compute Officer, Anthropic, June 30, 2026, available at www.politico.com/f/?id=0000019f-1b10-d591-a5ff-5b5d536a0000#:~:text=Fable%20models.,circumstances%20change%20or%20should%20Anthropic.

PROOFS

I.6.2 The Opacity, Unpredictability, and Non-replicability of AI Model Outputs

The power of AI companies is compounded by the nature of the technology, the opacity of which presents unprecedented challenges related to interpreting how AI models perform. Here, too, we see impacts both for politics and for political science research. Artificial intelligence presents novel difficulties for governance of a fast-moving, opaque technology. It also poses unique challenges to the practice of political science, which requires replicability and interpretability.

With respect to governance and regulation, as suggested earlier, AI's unpredictability and opacity present challenges not seen with earlier technology. Nuclear technology, to which AI is often compared, was and is inspectable by governments. Whether for weapons or energy, officials inside and outside of government understand how the technology works and can regulate it accordingly. For AI, governments can test and red-team models, but auditing standards (as discussed more fully further) are necessarily incomplete with respect to the panoply of AI use cases (especially outside existential risks), slowly evolving, and not yet standardized. As a result, it seems literally impossible for evaluators – inside or outside government – to predict and assess model performance according to normal modes of governance applied to other technologies, like cars or planes. Short-staffed on relevant AI expertise, governments are, therefore, often dependent on representations made by model developers. Worse still, given the aggressive competition in the AI ecosystem, new models (or significant updates) are often released without warning and without adequate time to evaluate their performance or impacts (although it remains to be seen if this will continue to hold in the post-Mythos world). To be sure, some of these dynamics can be seen in other sectors of public administration, like drug development, financial products, or software, but, none of these settings present general purpose technologies with similarly far-reaching implications.

The challenges for political scientists are equally serious. Using AI for research can be fundamentally different from analyzing a large data set or conducting randomized controlled experiments. Generative AI models are, at their core, probabilistic models, raising a question of how exactly we should think about the sensitivity or robustness of these models. Conceptually, we understand what we want to know here: How much do results change based on how we interact with these models? But probabilistic models should not be expected simply to produce the exact same output even when queried with the same prompt. There are some methods for making model outputs less variable (see, e.g., the discussion of "temperature" in Chapter 10), but some degree of variance in output is understood to lead to better performance. Thus generative AI introduces a trade-off between model performance and reproducibility, a characteristic not normally associated with classical conceptions of

PROOFS

robustness testing. Depending on the nature of the study, moreover, maximizing consistency to ensure replicability may undermine the benefits of using AI in the first place.

Relatedly, how does one ensure that AI-augmented research is replicable if the generative models are (1) probabilistic and (2) changing all the time? We are encouraged by recent efforts by some of the authors in this volume to establish research practices and norms to deal with these challenges, at least in part. "Guide-LLM" responds to some of these challenges by attempting to establish standards of transparency surrounding LLM usage in research.[11] The categories of reporting the authors included in the checklist are as follows:

- **Scope of LLM Use**
- **Model/System Details**
- **Prompts**
- **Data Inputs and Privacy**
- **Validation and Interpretation**
- **Guidelines for Reproducibility**
- **Competing Interests**

with an additional list of optional reporting:

- **Justification for LLM Choice**
- **Rationale for Prompt Design**
- **Comparison Against Other Methods/LLMs**
- **Training Data Leakage Risks Addressed**
- **Risk of Bias or Systematic Differences Affecting Conclusions**
- **Conversation Transcripts**
- **Ethical Implications of the Research**
- **Computational Resources**

Although this may seem cumbersome, we are encouraged by the ease with which agentic AI itself can likely be deployed to develop appropriate log files to allow for this sort of transparency (or even to automate completion of the checklist). More generally, our hope is that, with this volume, we might begin to push the discipline into establishing these or similar norms for political science research employing AI, which may very well emerge as a substantial category of research in the field in short order.

I.6.3 Addressing Bias in AI Models and Related Research

Several of the chapters in this volume emphasize the dangers posed by bias in AI models. Unsurprisingly, the chapters relating to gender and racial politics

[11] Feuerriegel et al. (2026).

PROOFS

do so extensively, but so do most of the other chapters in one or another respect. The chapters not only admonish users to be aware of the biases in these models and urge model developers to be transparent about these biases, but they also emphasize that bias in AI models is inevitable and baked into the process of training. Indeed, for both the politics of AI and related research, this volume represents a clarion call to recognize that bias is a feature of AI, not merely a bug.

The literature on AI bias is quite extensive, even if scholars ardently debate the appropriate ways to measure bias and which biases are politically consequential or worth addressing (Wang et al., 2025). On the one hand, famous examples of AI's racial bias, such as the COMPAS recidivism algorithm deployed in the criminal justice system,[12] have highlighted the way that AI can be deployed in real world settings with racially discriminatory results. On the other hand, we have the notorious examples identified by critics of "woke AI," such as Google's early image generator tools that led to query responses that depicted black World War II German soldiers and a racially diverse set of Popes or Founding Fathers for the US Constitution. Bias can creep into AI systems at every stage of their development – from the selection of data for training models to fine-tuning and to guardrails and filters that constrain an application's performance. Nevertheless, as the authors in the chapters that follow point out, a model that accurately represents the universe of content included in its training data will likely reflect the bias of the training data in its outputs.

Addressing AI bias is near the top of "political" considerations as policymakers within both government and the companies seek to erect guard rails against one or another bias. As AI tools roll out in arenas with well-developed antidiscrimination regimes, such as employment, housing, and education, existing laws need to be adapted to a world with human decisionmakers less "in the loop." Some jurisdictions, such as Colorado, which passed in 2024 the Colorado Anti-Discrimination in AI Law, have enacted broader laws that regulate AI as a technology to prevent discrimination against identified groups. Subsequently, the December 11, 2025, Executive Order, "Ensuring a National Policy Framework for Artificial Intelligence,"[13] specifically mentions the Colorado law, as an example of state laws that "are increasingly responsible for requiring entities to embed ideological bias within models." Thus, legal attempts to counteract bias are themselves accused of injecting other forms of bias – such is the politics of AI regulation in 2026 in the United States.

Of course, these concerns are not limited to the United States. The EU AI Act specifically requires high-risk AI systems to examine "possible biases that are likely to affect the health and safety of persons, have a negative impact

[12] Angwin et al. (2016).
[13] Executive Office of the President (2025).

PROOFS

on fundamental rights or lead to discrimination."[14] At the other extreme, Chinese models, such as DeepSeek, appear to have been adjusted in the post-training phase of model development to ensure they refuse to answer (or give regime-favoring responses) to some politically sensitive questions.[15] However, even without these kinds of post-training adjustments, if training data in certain languages tends to come disproportionately from government or government-approved sources, those political biases will likely "emerge" in model outputs, even in Western models (Waight et al., 2026). In addition to recognizing the underrepresentation of certain groups in training data that can lead to biased outputs, researchers need to be aware of the political realities that affect the very production of text in the first place that is used in model training. Social scientists, therefore, have an important role to play in understanding the contexts in which institutions affect the production of training data, and their concomitant consequences on the outputs of AI models.

Finally, beyond the question of bias itself lies the issue of *perceptions* of bias in the outputs of AI models. Much interesting research remains to be done on whether people tend to see AI-generated news as inherently suspicious or, conversely, more trustworthy than traditional sources of news. Surely, there are important moderating factors to be considered here, such as the demographic characteristics or ideological proclivities of the individuals perceiving bias (or a lack thereof) in AI output. Perceptions of bias also may vary depending on the subject of the query and the related output. However, both for AI policy and for political science research, public perceptions of the bias of AI systems may be as important as the actual bias baked into the models or the training data.

I.6.4 Benchmarks and Evaluations

One of the challenges to addressing biases in AI models, as well as models' misalignment or poor performance, is the absence of agreed-upon metrics for evaluating AI systems. Any "problem" identified in performance assumes some judgment as to how a properly functioning system would behave and some method to detect when and how far a model strays from the ideal. The evaluation science of AI models has become a field unto itself, with potentially high stakes both for the development of models and the study of AI systems.

To some extent, the articulation of proper benchmarks is both the most minimal and most critical role that policy can play in shaping the future development of AI. Whether the area of concern is bioweapon development, the ubiquity of hallucinations, or the likelihood that a companion bot leads a vulnerable individual to commit suicide, judgments that a model erred

[14] EU AI Act, Article 10(2)(f).
[15] Li (2025).

PROOFS

assume some ground truth or agreement as to what proper performance entails. Even if we could agree on how a perfect model might perform, policymakers must decide how much error is tolerable, given the countervailing benefits of a given AI application. For example, autonomous vehicles will end up causing fatal accidents. But how much risk should be tolerated, especially given that human drivers may be equally or more dangerous on the roads? These decisions are inevitably political ones, with cost-benefit analyses that are not necessarily unique to AI. Yet anxiety about the rapid AI rollout and residual fears from the last decade's battles over technology and politics may lead some to expect more from AI systems than we do for their human counterparts.

Gaming of benchmarks has also now become a frequent critique lodged at any new model. Whether the metric of choice relates to general knowledge, mathematics, reasoning, coding, safety, or any number of other performance measures, once the "test" is well known, model developers can "teach to the test." The establishment of consistent benchmarks that can measure, over time, relative performance of different models has become both essential and bedeviling for anyone who cares about AI policy and governance. Indeed, for all the hyperventilation over the specter of imminent development of AGI (Artificial General Intelligence), agreement over what constitutes AGI has remained elusive.

We see the same kinds of problems in political science research. When evaluating the performance or propriety of AI tools for a given project, we must ask – as with the autonomous vehicle example previously discussed – "compared to what?" How do we know when an AI model or tool performs "well" for the researcher? For example, as referenced throughout this volume, one of the most important use cases of AI to date in political science has been for classification tasks, such as categorizing text as about politics or not, reflecting an ideological position or not, being related to a particular topic area, and so on. Instinctively, human coding holds strong appeal as the gold standard against which AI must be benchmarked. Yet we know human coding for classification tasks is fundamentally inconsistent – otherwise we would not report intercoder reliability, as all human coders would be generating the same classifications. So if humans are flawed at classification tasks, is it possible that machines could do a better job? We already know that machines can perform classification tasks at a speed and scale that far outstrips humans. However, if we recognize the inherent shortcomings of human performance on similar tasks, against what benchmark should we evaluate AI performance? And how might such benchmarks differ across objective (e.g., is there a flag in the picture?) versus subjective (e.g., does the image convey empathy?) classification tasks? If machines are simply probabilistic models, though, what does it even mean for a machine to be "better" than a human at classification tasks that do not have a simple ground truth, such as classifying content as liberal or conservative or threatening or not?

PROOFS

And even if it is possible to answer this question, once we conclude that machines are "better" at classification tasks than humans, then how can we validate any further improvements in classification once we have lost human coding as the gold standard? And if we no longer have gold standards for validation of new methods, what does this ultimately mean about the reproducibility of research?

Versions of these questions permeate many of the chapters of this volume, but perhaps none more so than the chapter on AI and public opinion, which presents this dilemma at its starkest. As political scientists, we have enjoyed decades of methodological innovation based on the indisputable premise that simply asking a bunch of people their opinions is a fundamentally flawed way to measure public opinion. Statistical innovation after innovation has been aimed at figuring out how to better infer aggregate public opinion out of survey questions, but at the end of the day, even the best methods just do a better job of quantifying uncertainty. How then ought we to benchmark "silicon samples" of AI agents in an attempt to use these samples for measuring public opinion? If they match (flawed) human measures of public opinion? If they are subject to the same sources of error (e.g., question ordering effects) as humans? If they get the macro-level estimates of public opinion "right" even while getting micro-level measures "wrong"? All of these questions, and many more, require not just objective measures of whether benchmarks are being met but also subjective judgments of what should be measured – to say nothing of the normative judgment of whether human public opinion can or should even be measured using nonhuman agents.

1.7 CONCLUSION

It is difficult to describe the anxiety and excitement surrounding the editing of a book on AI and politics in 2026. As with evaluating the significance of AI, in general, none of the historical analogs quite captures the moment. Is AI like computers or the internet, turbocharging human capabilities and economic innovation? Or is it like television or even the printing press, which revolutionized public communication and education, with destabilizing effects on existing power arrangements? Is it akin to nuclear technology, as it holds a promise of significant human benefits alongside potential catastrophe? Or is it more like electricity or the steam engine, a general purpose technology with multitudinous applications that spur on an industrial revolution with accompanying social dislocation? Might AI represent an even more fundamental change, like the creation of the alphabet or discovery of agriculture, which represented stages in the early evolution of the human condition? Each of these analogies offers some insight into a dimension of this new technology, but none can fully capture the potential impacts on existing institutions, let alone the study of them.

PROOFS

Even if, as all the authors in this volume admit, it is difficult to predict the exact trajectory of this technology, we need to be grappling now with its implications for governance and the study of politics. It may be beyond our capabilities to assess the full range of political implications of AI, but political scientists have an obligation, we think, to track its development and to evaluate how political actors incorporate AI into familiar dynamics that political scientists study. Indeed, we also think that the discipline has an important role to play in providing the analysis that can help steer technological development and its governance toward socially productive ends.

In this sense, the chapters that follow represent a call for academics to join in the battle both to understand and to participate in the revolution that is underway. The challenges and opportunities this new technology poses for the pursuit of knowledge occur at a time when the very notion of human expertise and authority, across any number of dimensions, is being questioned. As politics remains a central domain and apparatus for the allocation of power in *human* society, those of us who study the impact of AI on political power need to use our minds and voices, even if inevitably augmented by the technology itself, to make sense of the new world this technology is creating.

1.8 EPILOGUE

On January 2, 2026, or approximately six months into the process of writing the chapters for this volume, Stanford political scientist Andrew Hall posted on social media that "Claude Code and its ilk are coming for the study of politics like a freight train."[16] The following day he posted online an empirical political science paper (replicating and extending one of his own prior papers) where the data collection, data analysis, and writing of the paper was all done entirely by Claude Code – in a matter of hours.[17] In the ensuing months, the freight train has gained speed as some political scientists have begun transforming their research pipelines with AI agents and even newer research-forward agentic tools such as Anthropic's Claude Science or Google Deep Mind's Co-Scientist. At this early stage of deployment, we can only speculate about what these developments will mean not just for political science research but for the very academic institutions that have long sustained that research.[18]

It no longer seems debatable, however, that agentic AI will turbo-charge some forms of political science research. It should also improve our ability to produce higher quality research thanks to the avalanche of review tools we are now able to throw at our work in progress. In addition to being able to use agentic AI to design research, draft pre-analysis plans, collect data, and

[16] Hall (2026a).
[17] Hall (2026b).
[18] Messing and Tucker (2026).

PROOFS

conduct statistical analyses, we can use it to check the robustness of those analyses, draft literature reviews, and compile entire papers. We also now have the ability to use AI agents to simultaneously check our code for errors, check our writing for style, accuracy, and within-paper consistency, and check the validity of our causal claims, the fidelity of our analysis to pre-analysis plans, our bibliography references, and much more. And while not every recommended change to a paper suggested by an AI agent is going to be a good one, the sheer quantity of feedback we can now receive on our work in mere hours (minutes) is stunning.[19]

If we accept the assumption that AI can accelerate the production of high-quality research, though, then an almost limitless set of questions emerges for political science as a discipline – and as a set of academic institutions designed to support research. We note in particular the following four issues:

1. What happens to our journals and our existing system of peer review as the time necessary to complete a sufficiently high-quality paper to submit to a refereed journal drops from months or years to weeks or days?
2. How might scholars *consume* new research, already a challenging task, if the speed at which it is produced continues to accelerate?
3. How might training of students evolve when so much of the work for which we previously employed research assistants – both undergraduate and graduate – can now be done much faster and cheaper by agentic AI?
4. Will AI tools that accelerate some forms of research more than others ultimately end up privileging the former over the latter? Might decisions about hiring and promotion end up privileging the type of work that can be accelerated by AI, or, conversely, doing the opposite by rewarding research that can not be replaced by AI (e.g., field work)?

Taken together, it is possible to imagine a future where largely AI-produced papers are reviewed by AI agents and then, once posted or published, are read and summarized by other AI agents for academics to peruse before being passed to yet another set of AI agents to generate ideas for new research papers as the cycle begins anew. It is a possibility that begs the question: Where exactly will we, as human political scientists, ultimately fit into this process?

We have not yet reached – and indeed may never reach – this version of a future for our profession where humans are increasingly crowded out of the research process by AI agents. But we are at the point where we at least need to begin thinking about it. Political scientists are not immune to the transformative potential that AI portends for other sectors of the economy. Like everyone else, we confront a professional future that could be both enhanced and undermined by the rise of artificial intelligence.

*****

[19] For just one excellent example, see Bäckman (2025).

PROOFS

REFERENCES

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine bias*. ProPublica. www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Accessed: April 21, 2026).

Bäckman, C. (2025). *AI-research-feedback* [Computer software]. GitHub. https://github.com/claesbackman/AI-research-feedback (Accessed: April 21, 2026).

Capstick, E. (2025). Public opinion. In *AI index report 2025* (Chap. 8). Stanford Institute for Human-Centered Artificial Intelligence. https://hai.stanford.edu/assets/files/hai_ai-index-report-2025_chapter8_final.pdf (Accessed: April 21, 2026).

Diamond, L. (2010). Liberation technology. *Journal of Democracy*, *21*(3), 69–83.

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council (EU AI Act). *Official Journal of the European Union*, *L 2024/1689*. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng (Accessed: April 21, 2026).

Executive Office of the President. (2025, December 11). *Ensuring a National Policy Framework for Artificial Intelligence* [Executive Order]. The White House. https://tinyurl.com/mu94hjy8 (Accessed: April 21, 2026).

Feuerriegel, S., Barrie, C., Crockett, M. J., et al. (2026). A reporting checklist for large language models in behavioural science. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-026-02492-7. https://llm-checklist.com/.

Hall, A. B. [@ahall_research]. (2026a, January 3). Claude code for research [Post]. X. https://x.com/ahall_research/status/2007221974947508303 (Accessed: April 21, 2026).

Hall, A. B. [@ahall_research]. (2026b, January 4). Proof of Claude Code for empirical polisci paper [Post]. X. https://x.com/ahall_research/status/2007603340939800664 (Accessed: April 21, 2026).

Li, J. (2025, January 31). What questions will China's DeepSeek not answer? *Deutsche Welle*. www.dw.com/en/what-questions-will-chinas-deepseek-not-answer/a-71470843 (Accessed: April 21, 2026).

Messing, S., & Tucker, J. A. (2026, March 3). *The train has left the station: Agentic AI and the future of social science research*. Brookings Institution. www.brookings.edu/articles/the-train-has-left-the-station-agentic-ai-and-the-future-of-social-science-research/ (Accessed: April 21, 2026).

Munger, K., Bakker, B. N., Berinsky, A. J., et al. (2026). *Peer review 2027: Scenarios for academic publishing in the age of AI* [Preprint]. SocArXiv. https://osf.io/preprints/socarxiv/594zj_v1 (Accessed: April 21, 2026).

Munger, K. (2023). Temporal validity as meta-science. *Research & Politics*, *10*(3). https://doi.org/10.1177/20531680231187271 (Accessed: April 21, 2026).

Persily, N., & Tucker, J. A. (eds.). (2020). *Social media and democracy: The state of the field, prospects for reform*. Cambridge University Press.

Quinnipiac University. (2026, March 30). *The age of artificial intelligence: Americans' AI use increases while views on it sour* [Poll]. https://poll.qu.edu/images/polling/us/us03302026_uai053.pdf (Accessed: April 21, 2026).

Sanderson, Z., & Tucker, J. A. (2023, November 1). Beyond red-teaming: Facilitating user-based data donation to study generative AI. *Tech Policy Press*. https://tinyurl.com/yfhmkssf (Accessed: April 21, 2026).

PROOFS

Waight, H., Yang, E., Yuan, Y., Messing, S., Roberts, M., Stewart, B., & Tucker, J. (2026). State media control influences large language models. *Nature*. 655, 685–693, https://doi.org/10.1038/s41586-026-10506-7.

Wang, A., Phan, M., Ho, D., & Koyejo, S. (2025). *Fairness through difference awareness: Measuring desired group discrimination in LLMs*. arXiv. https://arxiv.org/pdf/2502.01926 (Accessed: April 21, 2026).

Wike, R., Silver, L., Fetterolf, J., et al. (2022, December 6). *Views of social media and its impacts on society in advanced economies*. Pew Research Center. https://tinyurl.com/383a6td9 (Accessed: April 21, 2026).

PROOFS

1

Artificial Intelligence and Democracy

Campaigns, Elections, Movements, and Deliberation

Bailey Flanigan, Florian Foos, Archon Fung,
and Charles Stewart III

PROOFS

Every major advance in information and general purpose technologies changes democracy because that technology frequently enables political actors, as well as economic and social actors, to do their work more effectively. Technology can enable insurgents to displace incumbents, as some resources and capabilities (e.g., hegemonic support from mainstream media actors) lose value and other capabilities (e.g., ability to command attention on social media) increase in political value. On the other hand, technological change can also enable incumbents to tighten their grip on power if new technologies amplify political capabilities according to previously existing asymmetries of resources and power. In this chapter, we look into the near future to venture some reasoned speculation regarding the ways in which the proliferation of some artificial intelligence (AI) technologies may change democratic practices.

"Artificial intelligence" and "democracy" denote two enormous domains of activity and social inquiry. We therefore limit the scope of these reflections in both domains. Within "AI," we focus on the political applications of large language and multi-modal models (LLMs and LMMs) and generative AI. For the most part, we disregard other parts of AI, such as robotics, expert systems, and the possibility of Artificial General Intelligence. "Democracy" is an even larger domain. We limit our consideration to four of its aspects: campaigns, election administration, social movements (and governments' efforts to control them), and citizen deliberation. Furthermore, we look into the relatively near term and do not address longer term possibilities, such as AI agents enabling each individual to become a super-citizen on the one hand[1] or, on the other hand, Foucauldian total domination through hyper-surveillance by authoritarian state actors.

[1] See Hall (2026).

Tucker_9781047774734_C001.indd 26 07-09-2026 19:27:37

1.1 ENHANCING POLITICAL CAPABILITIES: INTO THE NEAR FUTURE

Our approach to thinking about the effects of AI is inspired by the task-based conceptual framework that Daron Acemoglu and Pascual Restrepo developed for reasoning about the effects of technology on labor and production.[2] In that framework, a job is a bundle of tasks. Rather than trying to figure out whether a technology will eliminate (automate) a job, we should think about which tasks the technology performs without labor (substitution), which tasks are left untouched, and which tasks might be done better with a combination of labor and technology (augmentation). We reason about the tasks that can be done, or done better, with generative AI in political campaigns, election administration, social movements, and citizen deliberation.

But what does "done better" with generative AI mean? The core capability of generative AI is to facilitate the manipulation of information in the form of text, images, and videos through actions such as searching, summarizing, translating, recombining, and generating. Humans also do these things. But, as Schneier and Sanders write, generative AI may help humans do them better, or do them better without humans, in four ways: by increasing speed, scale, scope, and sophistication – or all of these.[3]

While enhancing the quality of election administration seems unambiguously good for democracy, increasing the effectiveness ("doing better") of the other three kinds of political activities – campaigns, social movements, and citizen deliberation – could be good or bad for democracy. For example, enhancing the capabilities of political campaigns to win by misleading or mobilizing citizens' darker impulses may harm the quality of democracy. Amplifying the capabilities of ethnonationalist or illiberal social movements may also harm democracy.

1.2 POLITICAL CAMPAIGNS

Generative AI is a new technology, and its use by political campaigns is in its infancy. Nevertheless, political campaigns in countries as diverse as India, South Africa, Mexico, the United Kingdom, and the United States already utilize LLMs.[4]

People who work in political campaigns devote enormous amounts of labor to producing campaign materials for email, postal, and video platforms, fundraising proposals, and other written and video communication. Generative AI augments capabilities of campaign workers to produce a greater variety of potentially higher quality text, audio, and visual materials much more quickly. These capabilities are especially valuable to down-ballot, resource-strapped campaigns. For example, Democratic strategist Mike Nellis has founded

[2] Acemoglu and Restrepo (2019).
[3] Schneier and Sanders (2025).
[4] Foos (2024).

PROOFS

Quiller, a start-up that harnesses generative AI to produce drafts of fundraising emails for campaign workers to tailor and improve.[5]

Generative AI already features high-quality, cheap, and speedy voice and text language translation capabilities. In the 2024 Indian elections, Prime Minister Narendra Modi[6] used a generative AI tool called Bhashini to translate his campaign speeches[7] – delivered into Hindi – Tamil and other regional languages in real time. When he was mayor of New York City, Eric Adams used a translation tool from ElevenLabs to send messages to some four million people through the city's robocall system in languages he does not speak, including Spanish, Yiddish, Mandarin, Cantonese, and Haitian Creole.[8] Critics accused Adams of using "deepfakes" through these calls because he did not disclose that it was not actually him speaking, while the mayor responded that it was an effective way to reach the city's diverse population. Partially in response to this episode, the New York State Legislature enacted legislation that requires that any political communication in the state that uses AI disclose that the message has been manipulated by AI.[9]

While generative AI will continue to increase the efficiency of such "back-office" tasks as fundraising and language translation, it may have a more profound impact on political campaigns by increasing the scope and quality of a campaign's interactions with voters, relating to targeting, listening, persuasion, and mobilization. At the positive end of outcomes, generative AI can facilitate dialogue between campaigns and voters and make campaigns more responsive to voters' priorities.

1.2.1 Message Development

In the domain of targeted messaging, political consultants are developing technologies of "synthetic audiences," at large and small scale, that enable politicians to conduct polls and focus groups much more cheaply and to test political messages and discover political values with greater speed, at scale, and with greater scope. Adjacent to the politics industry, *The Times* (of London) reported in 2025 that it engaged a company called Electric Twin to create a synthetic audience of the British media public using the newspaper's database of several hundred thousand actual subscribers.[10] It used this model and synthetic audience to test marketing and communication strategies to, evidently, very good effect. Companies are starting to advertise similar synthetic audiences, focus groups, and polling services to political clients.

[5] Samuels (2023).
[6] Sakunia (2024).
[7] Banerjee et al. (2024).
[8] Guglielmo (2023).
[9] New York State Senate (2023).
[10] Peterson (2025).

PROOFS

1.2.2 Mobilization

Two fundamental aims of political campaigns are to (1) mobilize people who are already inclined to support their candidates and (2) persuade those who are not yet on board to become supporters. Get-Out-The-Vote (GOTV) campaigns are essential to a modern ground game. However, some GOTV contact modes that have been shown to be effective at increasing turnout, such as door-to-door canvassing and telephone calls,[11] are also difficult to scale up. Based on a meta-analysis of more than 700 GOTV interventions, Foos et al. find that the larger the sample size of a GOTV campaign, the smaller the size of its reported effect.[12] This association holds for interventions that contain an important human element in them, door-to-door canvassing and telephone calls, but not for GOTV interventions that are easy to scale up, such as mail shots, SMS text messages, and social media messages, making it unlikely that the association stems predominantly from publication bias in smaller studies. The authors attribute this smaller effect size in larger GOTV interventions that rely on humans for delivery to the well-known organizational challenges of scaling up social interventions, including the recruitment, training, and monitoring of human volunteers and workers, which are required to deliver an intervention at scale and at consistent quality.[13] While human canvassers will be difficult to replace at the doorstep, randomized trials have shown that AI chatbots are effective at mobilizing voters.[14] Already, candidates' phone calls can be delivered in a conversational manner by AI voice agents, substituting for human telephone canvassers and enabling campaigns to contact voters at scale via telephone calls.[15] It is not hard to imagine that those calls will be as effective as calls by humans and more effective than robot calls of the past. Another GOTV mode that can benefit from the involvement of AI chatbots is SMS text messaging. For a long time, campaigns sent one shot text messages to voters to remind them to vote; but, more recently campaigns have introduced interactive elements to those text exchanges, which now predominantly take place via messenger apps. Artificial intelligence chatbots are likely to replace human volunteers in interactive text messaging; that shift is already evident in other industries for many customer service interactions.

1.2.2.1 Persuasion

Beyond mobilization, multiple recent studies have found that conversations with LLM-powered chatbots can be politically persuasive – even more persuasive than traditional (non-AI) methods of political persuasion. Lin et al. (2025) conducted an experiment in which participants engaged in conversation with

[11] Green and Gerber (2019); Issenberg (2013).
[12] Foos (n.d.).
[13] List (2022); Al-Ubaydli et al. (2019).
[14] Mann (2021).
[15] Greschler (2024).

PROOFS

a chatbot that was instructed to persuade them to support a political candidate (e.g., Donald Trump or Kamala Harris) or a ballot measure (legalizing psychedelic drugs in Massachusetts).[16] They found that those exposed to the chatbot supported the bot's candidates and positions more than they did in the control conditions, at quite significant levels. The researchers conducted the experiment for multiple contexts: 2024 national elections in the United States, Canada, and Poland, as well as a ballot campaign in Massachusetts. Offer-Westort et al. (2026) present suggestive evidence from a randomized trial, replicating the intervention in Kalla and Broockman (2016), that AI chatbots integrated into Facebook Messenger were effective at reducing exclusionary attitudes among users.[17]

In a recent study published in *Science*, Hackenburg et al. (2025) tested the political persuasiveness of nineteen different LLMs.[18] Participants were assigned to chat with an LLM that was instructed to persuade them on an issue for two to ten rounds, while control participants read an LLM-generated static message. Both the static message and multi-round conversations produced statistically significant persuasion effects. Interestingly, large LLMs (more expensive) were more effective than smaller ones. However, "training" LLMs on conversations that were more persuasive increased their effectiveness more than model size, suggesting that it may be possible to create low-cost persuasive LLMs that are accessible to less well-resourced political campaigns.

1.2.3 Listening and Dialogue

Most political campaigns operate in a one-to-many, one-way mode: A candidate broadcasts their appeals to a mass audience of potential supporters and swing voters. Potential supporters sometimes have their say in smaller venues, such as house parties and town halls, and politicians also try to glean what voters want through surveys and focus groups. Over the last decade, listening has become an important task for campaigns in order to poll individual voters on political issues and to collect personal contact data. Listening campaigns involve human canvassers visiting voters at the doorstep or calling them on the phone. Listening campaigns often accompany the rollout of a campaign and have been applied widely, for instance, by Emmanuel Macron's 2017 presidential campaign[19] and by Hillary Clinton in her 2016 presidential campaign.[20] While information still flows from the candidate to the audience in most campaigns, generative AI applications may alter this dynamic by *increasing* the flow of information from constituents to politicians through listening and

[16] Lin et al. (2025).
[17] Offer-Westort et al. (2026).
[18] Hackenburg et al. (2025).
[19] Reuters (2016).
[20] For this Hillary Clinton listening tour, few hugs but much note-taking (2015).

PROOFS

dialogue. Artificial intelligence can augment a political campaign's abilities on the listening side in two ways: It can help with the systematic analysis of text and voice responses, and it can conduct conversations at scale. While listening campaigns collect a lot of texts from citizens, they have struggled in the past to analyze these open text responses effectively. Large language models are easily able to identify issues, positions, and tone of open text or voice answers and provide campaigns with real-time means to aggregate voter sentiment and issue priorities.

Large language models combined with online political discussion platforms and apps also have the potential to enable campaigns and politicians to "listen" to citizens and constituents at scale, with increased sophistication, by engaging almost unlimited numbers of citizens in many-round discussions, identifying the most common themes and points of agreement and locating issues of majority or super-majority agreement.

"AI Steve" may have been the first LLM-powered politician to run for office.[21] "AI Steve" is the digital avatar of Steve Endacott, who ran as an independent seeking to represent the Brighton Pavilion constituency in the United Kingdom Parliament in the 2024 General Election. Electoral authorities made clear that, if he won, Endacott – not "AI Steve" – would be the member of parliament. But Endacott proposed that the candidate, and the legislator, if elected, would be a cyborg combination of Endacott and "AI Steve." "AI Steve" was directed to have conversations on policy issues with people in Endacott's constituency. It was capable of carrying on 10,000 simultaneous conversations and was available 24/7. The avatar would record and analyze all conversations, identify positions with majority support, and then Endacott would advocate for those positions. In the July 2024 election, "AI Steve" lost, having received only 179 votes (less than 1 percent of the total votes cast).[22] But, as this experience illustrates, it seems likely that different kinds of human + AI chimeras will become more common in politics.

A more prominent recent experience of a human politician + AI chimera comes from Pakistan. In the election of early 2024, candidates who were aligned with the Pakistan Tehreek-e-Insaf (or PTI) party won the most votes for seats in parliament. Ruling authorities had imprisoned PTI leader Imran Khan since 2023. Khan's supporters used AI technologies to construct a video of Khan delivering a victory speech, and they disseminated it widely.[23] They also campaigned on behalf of Khan, using AI to construct speeches from notes that Khan wrote from prison.

Dialog as a political activity is closely related to listening. Dialog represents a two-way exchange between politicians and constituents and may help politicians and constituents to be more informed about one another and increase the

[21] Elliott (2024).
[22] UK Parliament (n.d.).
[23] Imran Khan uses A.I. to give victory speech in Pakistan (2024).

PROOFS

confidence and trust that constituents have in their representatives.[24] But for candidates and representatives, is time-consuming and expensive – so expensive that scholars who study money and politics have argued that campaign contributions "buy" access[25] or the opportunity to engage in dialogue with one's political representative.

Applications of generative AI have the potential to decrease dramatically the cost of dialogue by enabling constituents to engage in conversations with virtual LLM-powered avatars that are capable of answering questions about the politician and their positions and perspectives. Takahiro Anno used these technologies in his bid to become Tokyo's mayor (known as governor). Schneier and Sanders write that "Anno, a software engineer by trade, invented new political technologies and leveraged AI to amplify his individual capacity to listen. He used an AI avatar trained on his political manifesto to respond to 8,600 questions from voters over a seventeen-day continuous livestream."[26] Though he was an outsider candidate and lost the mayoral race in 2024, he ran for a seat in the upper house of Japan's National Diet in 2025 and won. In the course of his campaigns, he founded a new political party called the "Future Party" that promises to utilize AI and other technologies to usher in new methods of digital democracy.

Dean Phillips, a long-shot candidate for US president in 2024, also utilized an LLM-powered avatar; he called it "Dean.Bot." His supporters commissioned a start-up called Delphi to build the bot using ChatGPT. They reportedly trained the Dean.Bot on the candidate's speeches, podcasts, and interviews – totaling some 240,000 words.[27] Users could ask "Dean.Bot" questions verbally and receive spoken responses in a voice that sounded like Phillips, but "Dean.Bot" was a very short-lived effort. It went live in early January 2024, but OpenAI – the company that runs ChatGPT – shut the effort down by late January 2024 because its terms of use prohibit political campaigning.[28]

1.2.4 Citizens' Private Use of AI Chatbots

One under-researched area is citizens' private use of AI chatbots to obtain political information. Citizens are increasingly using AI chatbots in their everyday lives to obtain information about everyday issues, including politics and elections. These interactions with AI chatbots include users fact-checking political claims that friends are making in conversation, querying the political positions of candidates and parties, and asking about political events and

[24] Neblo et al. (2018).
[25] Kalla and Broockman (2016).
[26] Schneier and Sanders (2026).
[27] Silicon Valley insiders are trying to unseat Biden with help from AI (2024).
[28] OpenAI bans developer of Dean Phillips bot (2024).

PROOFS

practical voting matters. As more citizens rely on chatbots and other forms of AI to search for political information, the effects on the quality of the information they receive, and thus the information environment for democracy, strike us as radically underdetermined. On one hand, AI can provide highly relevant information in responsive and accessible ways – more so than a web search or perhaps even newspaper and magazine articles. On the other hand, we do not know whether this shift will exacerbate the polarization of information spheres or whether it will aid or hamper truth-seeking efforts. Those future patterns depend in part on choices and behaviors of individuals, political leaders, AI companies, and others. Will individuals seek out biased information to confirm or disconfirm their views? Will the AI that companies build provide accurate information and will they address viewpoint diversity?[29] In the provision of information through AI, how will companies balance truth-seeking versus user engagement? For example, will profit motives lead companies to push information to individuals based upon how much campaigns, politicians, and economic actors are willing to pay? The answers to these questions will be determined not in the first instance by the arc of technological development but rather by political, economic, and regulatory decisions.

1.2.4.1 AI and Personal Data Regulation

The scale at which campaigns can use generative AI in their campaigning depends on the willingness, and the ability, of countries and other political entities to regulate campaigns' use of generative AI systems[30] and their ability to harvest and share citizens' personal data legally.[31] While the United States has so far abstained from regulating the political use of generative AI at the federal level and has comparatively permissive data privacy and data-sharing laws, the European Union has taken concrete steps to regulate the use of AI in elections via the 2024 EU AI Act. Classifying AI systems intended to influence voting behavior or electoral outcomes as high risk, the AI Act imposes conditions for AI-enabled campaign interventions. Generative AI outputs need to be marked as AI generated in a machine-readable format and, in some cases, need to be visibly labeled. Similar regulations, which require campaigns to disclose whether content was produced by generative AI and whether voters are interacting with AI systems, have, as of 2026, been adopted by twenty US states.[32] California has taken a particularly strict approach to regulating the use of AI in election campaigns. The European Union had already previously, in 2016, restricted organizations' ability to harvest and share personal data

[29] Westwood et al. (2025).
[30] Foos et al. (2026).
[31] Foos (2024).
[32] Orrick (n.d.).

PROOF

via the General Data Protection Regulation (GDPR), which regulates the collection, processing, and sharing of personal data. Obtaining personal contact details at scale, and particularly telephone numbers, is a necessary for AI-driven GOTV and persuasion tactics that use phone or SMS text messages. Richer data, such as individuals' consumption and online behavior, are also important for targeted and tailored persuasion efforts.[33] These regulations, hence, directly pertain to the scalability of GOTV interventions where AI-based systems are used to replace human volunteers or to micro-target voters. Artificial intelligence–based gains in scalability will therefore, at least to some extent, depend on favorable national and supranational AI and data regulation environments. There is evidence that voters are in favor of stricter AI regulation if presented with campaigns' use of generative AI that could be labeled as deceptive, but that they are unlikely to punish parties for using these tactics.[34] This speaks to the incentives that different types of political actors face when deciding whether or not to deploy generative AI legally in voter contact or illegally, for instance, via the use of deepfakes or by not labeling content in line with regulations. Besides the regulatory environment, as Foos (2024) argues, smaller parties that have no membership base and are less bound by established political norms and inertia should have a larger incentive to rely on AI-bots in their campaigning than mainstream parties.[35]

1.3 ELECTION ADMINISTRATION

Like many organizations whose work involves communicating, interacting with the public, and responding to requests for information, local election officials (LEOs) have the potential, with the use of AI, to operate more effectively, with less effort and fewer resources. This is especially important because election administration offices are often highly strapped for resources, particularly in the United States. At the same time, perhaps because many LEOs are operationally conservative, few have begun to utilize the novel technology of generative AI in any significant way. According to surveys conducted by the Brennan Center, just 8 percent of LEOs[36] had used AI in their offices by 2025 (up from 5 percent in 2024).[37] The most common uses were drafting social media content and press releases, creating graphics, and language translation.

Like people working on political campaigns, LEOs must produce many different kinds of materials to inform voters: educational materials about how to register to vote; where and when to vote; voter ballot guides about offices,

[33] Fung and Lessig (2023).
[34] Jungherr et al. (2026).
[35] Foos (2024).
[36] Brennan Center for Justice (2025a). Also see National Conference of State Legislatures (2026).
[37] Brennan Center for Justice (2025b).

PROOFS

candidates, and issues; "I voted" stickers and other civic "swag"; and materials about the process of validating and counting votes. The last is increasingly important for defending against claims of voter fraud and other election malfeasance. In addition, LEOs are responsible for producing training materials that translate state-mandated requirements about voting procedures for the hundreds of thousands of workers who staff polling places in each election. Generative AI can greatly speed the work of producing such materials, though election officials must be careful to edit and verify the accuracy of final versions lest they unwittingly propagate misinformation that disenfranchises voters or erodes confidence in the election machinery.

Generative AI could also be used to respond to the many routine requests that election officials receive. These include information requests from voters and other residents: Am I eligible to vote? How do I register to vote? Where should I be registered? How and where do I vote? How do I become a poll worker? Others, if they are concerned about the integrity of how the election process is conducted, seek information about the processes and personnel involved in election administration. And election officials report increases in records requests that have become a substantial burden in some jurisdictions.[38]

As in many other industries that have demanding customer support requirements, some election officials have created chatbots and other technologies that interact via voice, text, or website to respond to high-volume and frequent inquiries from citizens. Many jurisdictions offer web forms that enable residents to check on their voter registration status and correct errors. Idaho election officials, for example, worked with IBM to use its Watson technology to create and deploy a chatbot to respond to voter inquiries about, for example, where to vote or how to acquire an absentee ballot.[39] At this writing, publicly documented examples remain limited, and election offices that deploy AI chatbots appear to favor retrieval-augmented systems grounded in official materials.[40] Many election officials are concerned that LLMs will generate hallucinations and other errors.

In a 2024 Columbia University workshop, participants tested the voting information veracity of several unmodified commercial chatbots, including OpenAI's ChatGPT-4, Meta's Llama 2, Google's Gemini, Anthropic's Claude, and Mistral.[41] Participants rated half of the chatbot responses as inaccurate and 40 percent as harmful. For example, according to PBS News (2024), "In Nevada, where same-day voter registration has been allowed since 2019, four of the five chatbots tested wrongly asserted that voters would be blocked from registering to vote weeks before Election Day."

[38] Yoder and Tan (2024).
[39] Freed (2020).
[40] National Conference of State Legislatures (2026).
[41] PBS News (2024).

PROOFS

The current common wisdom in the election administration field is to refrain from using generative technologies to automate responses to requests for information and rely instead on rule-based systems and other such technologies.

In recent years, a different kind of information query has become more common in election administration: the public records request. Local election officials have fielded an increasing number of public records requests since the battles over the integrity of the 2020 US presidential election.[42] Whether these requests stem from misinformation, malicious intent to paralyze election machinery, or standard inquiries from academics and reporters, they consume enormous amounts of time and resources in some jurisdictions. The Center for Election Innovation and Research reports that

> … in Wake County, North Carolina, officials reported a sevenfold increase in the number of election-related records requests received from 2020 to 2022, leading the county to increase its budget to hire a new staff member dedicated to processing incoming requests. In 2022, officials in Maricopa County, Arizona, reported one request that required nearly half the election office's staff to spend four days sorting and scanning 20,000 documents … Local election officials in states like Florida and Michigan have reported spending anywhere from 25% to over 70% of their time just on processing public records requests in recent years. Officials in states like Virginia have complained that they have been unable to engage in routine tasks like voter outreach due to the volume of requests received. (Yoder & Tan, 2024, p. 2)

The proliferation of generative AI may exacerbate this problem by making it easier for citizens to make such burdensome records requests. On the other hand, LLMs may make those requests less burdensome by enabling election officials to respond more quickly and with less effort. Though we could not find cases of local election officials who have built and used generative AI to speed records requests, there are promising examples from other governments. For example, a team in the US State Department trained an AI on diplomatic documents in order to see whether it could accurately classify documents that should be declassified.[43] The classifier turned out to be highly accurate, and the team expanded the project to respond to Freedom of Information requests by identifying the documents that should be returned, based in part on similar requests in the past.

There are several high-risk areas of election administration that could be sped up through partial AI automation. But, at this stage in the development of generative AI – and other forms of AI – mistakes are harmful (e.g., could cause disenfranchisement). Such tools should be used experimentally and with human oversight. For example, many jurisdictions validate mail-in ballots by matching handwritten signatures. Many jurisdictions utilize Automatic

[42] Trump backers inundate election offices with requests for 2020 records (2022).
[43] Partnership for Public Service (n.d.).

PROOFS

Signature Verification, a pattern discriminating (not generative AI), as a first pass to flag questionable signatures for human assessment.[44]

Generative and other forms of AI have the potential to improve the operations of too-often under-resourced agencies of local election officials. Some of the tasks that election officials must do can be enhanced by consumer level generative AI – chatbots such as OpenAI's ChatGPT and Google's Gemini. These chatbots, for example, are well suited to producing drafts of press releases, informational brochures, and poll worker training materials. They are also already quite capable of translating election materials into different languages. We expect generative AI to proliferate in these domains. Other tasks require generative AI applications that are customized – "trained" – with specialized information and domain-specific materials. For example, chatbots that provide information to voters will need to accurately convey the specific rules and provisions for particular states, and even counties and cities. The spread of these generative AI applications depends in large part on supply-side questions: Will technology vendors develop capable applications of these technologies? That depends, in turn, on whether local election officials will have access to resources sufficient to purchase such applications and so incentivize technology firms to develop them.

From a larger vantage, well-functioning democracy depends on systems of election administration that are both trusted and trustworthy.[45] Generative AI may help to increase trustworthiness by increasing the efficiency, reliability, and service quality of these agencies. It may increase trust by enabling local election officials to communicate more effectively with voters and media audiences in their jurisdictions. But on the other side, skeptics and critics of election administrators also have access to generative AI tools that enhance their capabilities to communicate and persuade.

1.4 SOCIAL MOVEMENTS

Artificial intelligence creates great challenges and potential opportunities for civil society organizations and social movements. Civil society organizations are struggling to constrain and mitigate the negative effects of AI on many fronts: algorithmic bias, privacy violation and surveillance, harms to people and the environment that stem from AI training and the construction and operation of data centers, and so on. On the other hand, AI can help to make social and democracy movements more effective and, potentially, reduce power asymmetries between movement organizations and the larger, better resourced governments and private organizations against which they often struggle.

[44] Cortés et al. (2023); U.S. Election Assistance Commission (n.d.).
[45] Stewart (2022).

PROOFS

Like political campaigns, social movements depend upon building popular support and raising funds in order to grow and accomplish their goals. Public communication – in person, through mass media; through social media; and through spoken word, audio, images, and video – is a principal method to build such support. As in many other fields, people in social movements can use generative AI technologies to produce text, images, and videos with greater speed and sophistication.

Freddy Guevara is a democratic activist and a leader of the Venezuelan opposition in exile. He writes, "I use AI almost daily in my strategizing, activism, and research.… you don't need advanced skills – just access to a trusted LLM and basic knowledge for asking the proper questions and creating the right prompts."[46] Creating well-produced content – especially images and videos – can require professional services that are too expensive for many social movement organizations and activists (but readily available to governments and private corporations). But generative AI tools – such as DALL•E and Midjourney for images and Synthesia and Sora for video – make it possible to produce sophisticated content for much less money, in much less time.

An early defining moment in social movements' use of AI-generated content was the quick and somewhat controversial spread of an image called "All Eyes on Rafah" in 2024. On May 26, 2024, The Israeli military raided a large displacement camp, holding some one million people, in the city of Rafah in the southern Gaza Strip. The raid killed several dozen people.[47] In the immediate aftermath, eyewitnesses shared photos of the devastation on social media, but platforms' content moderation algorithms (especially Instagram) blocked the images because they were too graphic.

Soon after, an image now known as All Eyes on Rafah began to circulate on Instagram and other platforms. The AI-generated image depicted a sea of tents with large white tents spelling "All Eyes on Rafah." At its peak in late May 2024, the image had been shared between 47 million and 50 million Times on Instagram. It also spread widely on X and TikTok. Its spread was fueled in part by celebrities, such as Bella Hadid, Dua Lipa, Pedro Pascal, and Alia Bhatt (a major Bollywood star).[48] The image spurred several controversies. Two different Malaysian individuals claimed credit for creating the image.[49] Critics also complained that the image sanitized the war by omitting the blood and bodies in favor of clean rows of tents. The image also overshadowed, and perhaps thereby de-valued, the work of actual journalists who documented the attack.

The Centre for Responsible Union AI conducts training programs for hundreds of union staff and members in the United Kingdom, Europe, Australia,

[46] Guevara (2025).
[47] Burgess and Morris Grant (2024).
[48] Burga (2024); Dunworth (2024); ABP News Bureau (2024).
[49] Allyn (2024).

PROOFS

and the United States about how to utilize AI tools. In a survey of their program participants, members wanted AI largely to help them with back office tasks, the mundane rather than the revolutionary: writing, managing email and meetings, research and analysis, training, and member engagement.[50] For example, Belgian union ABVV-Metaal created an AI-backed process to help union staff write letters (all containing standard information) requesting leave for members.[51]

One emerging bespoke use of AI by labor unions – using generative AI, natural language processing, and rule-based systems – is to create chatbots running through apps and messaging platforms. These chatbots help union members understand collective bargaining agreements and their rights and prerogatives under law. They also provide a means to file grievances, to report misconduct or workplace violations, and to recruit new members. For example, the UK Public and Commercial Services Union (PCS) created a chatbot to help train and simulate recruiting conversations for existing members to use in recruiting new members.[52] A large UK teachers' union (NASUWT) created a chatbot trained on its large internal library to help members deal effectively with many different kinds of challenging workplace situations.[53]

In addition to crafting communications to amplify social movement framing of their narratives[54] and accelerating "back office" tasks, movement activists and organizations have utilized AI (both generative and rule-based) to help recruit supporters and engage them in political activities.

An important example of such engagement comes from the 2024 Venezuelan elections. In that context, government authorities could not be counted on to run a free and fair election and to provide ample information and opportunities for Venezuelans to participate. Responding to this hostile electoral environment, the democratic opposition movement sought to create in-effect a parallel election apparatus that would, among other things, inform voters about where and how to vote, enable members of the public to report election improprieties, and – amazingly – create a parallel vote-counting apparatus to challenge what turned out to be fraudulent official results.[55]

Venezuelan social movement and political opposition activists worked with technologists to develop several chatbots to provide information and engage people in the 2024 elections.[56] They developed Cheo, a natural language chatbot operating on top of Facebook and Instagram and then later WhatsApp. Cheo provided real-time information about polling locations – which can be difficult to ascertain in Venezuela because there are more than 3,000, often in people's

[50] Centre for Responsible Union AI. (n.d.-a).
[51] Centre for Responsible Union AI. (n.d.-b).
[52] Centre for Responsible Union AI. (n.d.-d).
[53] Centre for Responsible Union AI. (n.d.-c).
[54] Benford and Snow (2000).
[55] Schmidt et al. (2024).
[56] Picón (2025).

PROOFS

homes, and they can shift rapidly. Activists developed capacities for Cheo to provide poll worker training for individuals and to help them get certified. Some 158,000 users signed up and interacted with the poll-worker training bot.

Activists also developed a Toma El Control chatbot to enable observers to report incidents of suspected election malfeasance, complete with geocodes, and the ability to upload photographic and video documentation. Researcher and participant Isabella Picón reported that "Over 200,000 messages were exchanged, receiving 88,885 and sending 112,080 responses with instructions to address interference at the polls."

A third broad domain in which AI can boost the efforts of social movements is research. Activists engage in research for many purposes: to root out corruption; to develop better advocacy and campaign strategies; and to identify government abuses and rights violations. Artificial intelligence tools can augment the capabilities of human researchers and analysts in social movements by speeding up research activities and reach scales that are not possible for small organizations to achieve with human effort alone.

Some efforts to identify corruption, bias, and abuse involve analyzing reams (or gigabytes) of government records. The Global Investigative Journalism Network, for example, has developed a method to identify and expose official gender-based discrimination and violence. The method involves using ChatGPT to scan many thousands of pages of court records to identify suspicious cases and incidents that merit further investigation and can constitute broader patterns of abuse.[57]

In another example, Nathan Sanders developed tools that use generative AI to analyze oral and written testimony in state legislatures in the United States. He worked with an organization called Health Resources in Action to utilize the tool to analyze more than a thousand oral testimonies in state legislatures to track emergent legislation. He also used it to fact-check testimony and characterize the values and ideological orientations appearing in this vast body of public testimony.[58]

Will social movements be able to harness the potential of AI to increase the speed, scale, scope, and sophistication of their activities – and hence their effectiveness and power – in these and other ways? They face distinctive challenges when it comes to harnessing the emerging technologies of AI. While some of the tasks of activists can be enhanced through off-the-shelf consumer technologies like chatbots, other tasks require tailored tools – such as the large-scale analysis of government documents or bespoke chatbots to engage supporters in elections or protests.

Whereas political campaigns and election officials are likely to rely on commercial vendors to develop custom AI applications and tools, social movements may be more reliant on networks of movement-friendly developers who

[57] Ulrich (2025). Discussed in Schneier and Sanders (2025).
[58] Schneier and Sanders (2025) chapter 35 "Watching the Government."

PROOFS

provide open-source tools.[59] First, many social movements and civil society organizations face even tighter resource constraints than political campaigns or election administrators. Thus, an ecosystem of vendors providing high-quality tools that these organizations can afford is less likely to develop. Second, in the face of hostile governments with formidable surveillance capabilities, many movement organizations place a high premium on privacy and data security. These activists may not trust private sector vendors who sell applications and tools based on code that cannot be inspected, audited, or easily changed.

1.5 CITIZEN DELIBERATION

Practices of citizen deliberation are as old as democracy itself: from discourse in the Athenian agora to town meetings in New England and open meetings today. In the contemporary political context, many see increased citizen participation in politics and public affairs as a way to strengthen democratic practices at a time when trust and confidence in democratic institutions and political leaders are in steep decline. At the same time, public involvement beyond casting a vote has long faced important practical limits – around scale, participation, and inclusion. There is growing optimism that some of these limits may be alleviated by the careful use of AI.

This section illustrates how AI's role in citizen deliberation is best understood not as a single intervention, but as a diverse toolkit for addressing recurring bottlenecks in deliberative processes. Different AI technologies can play different roles at different points in the democratic pipeline: facilitating registration, summarizing data, translating legal terms and election language, clustering, or mediation. In doing so, they potentially improve some aspects of voter participation and enable entirely new deliberative formats, while introducing distinct tradeoffs. The key question is often not whether "AI helps deliberation" in general, but which AI tool is best suited to address a given need, how that tool shifts discretion and control within the process, and what forms of accountability remain.

We begin by outlining four recurring bottlenecks in citizen deliberation. Then, we present three case studies, which offer examples of how AI tools have been deployed toward alleviating some of these bottlenecks. Finally, we synthesize across these case studies to unpack the key opportunities and risks associated with using AI to support citizen participation in democracy.

1.5.1 Four Key Bottlenecks

One of the most widely discussed challenges faced by direct public involvement in the political process is scaling, that is, including many people as participants. In *After the Revolution*, Robert Dahl describes how a town meeting cannot

[59] Jie (2025).

PROOFS

remain both inclusive and participatory as the number of citizens attending any given opportunity grows. In many, if not most, public meetings, only a small fraction of attendees can speak, or else a routine meeting or special hearing runs impractically long. Either way, "the greater the numbers, the more the town meeting runs the risk of becoming unrepresentative."[60] In addition to these logistical tensions, a major barrier to scaling is the need for high-quality facilitation: face-to-face deliberations often feature trained facilitators to keep discussions focused and make sure that participants feel included and engaged. Many public meetings (town halls, public hearings) lack sophisticated facilitation and, consequently, suffer from excess conflict, polarization, and the lack of meaningful exchange. Of note, nontechnical solutions have already made important progress on the scale problem: Since the 1990s, organizations such as AmericaSpeaks[61] have created methods including the 21st Century Town Meeting[62] to conduct face-to-face public deliberation at scale. Its two-level deliberations – in which many groups of eight to ten people simultaneously discuss a single topic (e.g., a city's budget) – allow several thousand individuals to participate in a single deliberative event.

A second significant barrier to direct public deliberation is *sensemaking*: With so many ideas being exchanged and refined among many people, a core challenge is how to synthesize this information, suggest avenues of common ground, and aggregate ideas into intermediate or final conclusions. Traditionally, the First two tasks are done by participants and facilitators. For example, methods like the 21st Century Town Meeting rely on humans to read and analyze notes from deliberations in real time and spot emerging themes for subsequent deliberative consideration and group convergence. Similarly, aggregation is often done by asking participants to vote on resolutions or by polling their agreement with relevant statements. These methods can be dissatisfying because they can obscure the reasons why participants support the positions they do, how and why their views change over the course of public deliberations, and areas and opportunities for expanding agreement or minimizing excess conflict.

The third key bottleneck in citizen deliberation is *unequal participation and inclusion*. This relates to Dahl's worry about whether someone can meaningfully impact – and thus be meaningfully included in – a large conversation. It also encompasses many other issues, including limited and unequal participation rates in synchronous deliberative processes like citizens' assemblies,[63] and discursive and circumstantial differences (e.g., differences in language or lived experience) that mean that people are best able to contribute to the "deliberative sphere" in different deliberative formats.

[60] Dahl (1990).
[61] See article on Participedia (n.d.-b).
[62] See article on the 21st Century Town Meeting on Participedia (n.d.-a).
[63] Flanigan et al. (2021).

PROOFS

Deliberative forums also face a fourth bottleneck around *trust and perceived legitimacy*, both in the general concept of deliberative forums and in specific deliberative events. Public trust in deliberative forums is not automatic, and it does not appear to depend only on whether deliberation occurred; rather, it is shaped by factors such as who is perceived to have participated,[64] whether the process is understandable to non-participants, and whether decision-makers visibly respond to the resulting recommendations.[65] More broadly, the conditions under which deliberative processes generate trust or legitimacy gains remain only partly understood, especially as these processes are adapted across institutional settings. Theoretical legitimacy, deliberative quality, and positive political outcomes may be insufficient if the process and its outputs are not legible to participants, observers, and downstream decision-makers.

1.5.2 Three Case Studies

1.5.2.1 Case #1: vTaiwan and Pol.is

Audrey Tang – Taiwan's first Minister of Digital Affairs – has been perhaps the most prominent public official using digital tools to enable public deliberation and gather direct input from constituents. In the mid 2010s, Tang deployed the *Pol.is* platform as part of her *vTaiwan* initiative to engage Taiwan's population in policy deliberations. The Pol.is platform is similar to a wiki-survey: Participants can contribute policy positions in text form and indicate their agreement or disagreement with a series of others' policy positions. Then, the platform uses clustering techniques and visualizations to highlight regions of agreement and convergence based on these agree/disagree voting patterns.

The policy application of this technique that has received the most attention concerns the regulation of Uber and other ride-sharing services.[66] Visualization of initial positions revealed polarization, but ultimately the differing sides actually reached agreement on several core principles: fair liability insurance, transparent pricing, and no undercutting of market rates. These principles were incorporated into the government's regulation. Some observers have credited this participatory process with increasing support for, and perceived legitimacy of, the resulting policy.

While vTaiwan's original launch predated modern LLMs, recent deliberative processes in Taiwan increasingly incorporate these technologies. For example, the "Talk to the City" platform uses LLMs to aid in facilitation, synthesize open-ended input, identify areas of overlap, and map potential paths to convergence.[67]

[64] Pow et al. (2020).
[65] Germann et al. (2024).
[66] See Cushing Rodriguez (2023).
[67] See Tseng (n.d.).

PROOFS

1.5.2.2 Case #2: The Habermas Machine

A second case comes from researchers at Google DeepMind and multiple universities. They have proposed an AI-powered tool called the "Habermas Machine" that is designed to improve public deliberations on complex and controversial issues. This tool works as follows: Participants submit their individual views in text format on the issue at hand and, in each round, an LLM drafts a set of statements intended to reflect competing perspectives while maximizing broad endorsement. Participants then evaluate these statements (e.g., approve/disapprove and provide feedback), and the system iteratively revises them. Unlike Pol.is – which primarily maps opinion clusters from voting patterns – the Habermas Machine uses generative AI in the mediator role to synthesize and refine candidate common-ground statements.[68]

Researchers evaluated the Habermas Machine in a multi-experiment study in the United Kingdom, which included randomized online experimental tasks and a virtual deliberative process. In the deliberative process, 200 participants met for three weekly one-hour sessions and deliberated over nine contentious policy topics using the Habermas Machine. They found that the LLM-produced common-ground statements were often positively endorsed (somewhat more often than those from human mediators), and within-group agreement increased on multiple issues from pre- to post-deliberation. This study's findings suggest that LLMs could usefully assist with mediation and common-ground discovery in deliberation. On the other hand, some critics argue that these observed opinion changes may simply result from people's tendencies to defer to algorithms, so more investigation is warranted.[69]

1.5.2.3 Case #3: Digital Dialogues in Libya

A third example comes from Libya in 2020–21. After years of civil war in Libya, a team working with the United Nations sought to engage many Libyan citizens in a series of public deliberations about topics central to the country's transition to peace.[70] Working with Remesh AI, they ran large-scale "Digital Dialogues" on an online, mobile-accessible platform. Participants provided both open-ended responses and answers to multiple-choice prompts. The key technical contribution was an AI-assisted analysis pipeline: The system rapidly organized open-text input into themes, estimated support for proposals, and compared patterns across participant subgroups (e.g., minority-group membership). Though its implementation was different, the AI system's purpose here was similar to that of the Habermas Machine, contributing sensemaking across multiple kinds of information. These dialogues were conducted in local dialects, and they were broadcast on both social media and live TV so that the rest of the population could witness them.

[68] Tessler et al. (2024).
[69] Cohen and Kugelberg (2025).
[70] Irwin et al. (2021).

PROOFS

The team used this platform to conduct five digital dialogues between October 2020 and January 2021, with 1,000 Libyans participating in each round. Recruitment was handled by a local polling company. These dialogues addressed impacts of the war; militias and other fighters; economic issues; oil revenues; human rights; and elections. In one dialogue, participants proposed questions that they would like candidates for the new Government of National Unity (GNU) to publicly address. Organizers write that "the questions collected through the Dialogue were later put to the GNU candidates who answered them on live television. These events achieved social media audiences of 1.7 million, a third of the Libyan population." As such, organizers present these Digital Dialogues as having informed – and potentially strengthened the legitimacy of – the process of forming Libya's interim government in 2021.

1.5.3 Zooming Out: Opportunities and Challenges

In the case studies discussed previously, we saw AI helping increase the *scale* of participation by connecting people across long distances, allowing lower-cost online participation, reducing language barriers, and reducing the need for extensively trained human facilitators. In Taiwan, tools such as Pol.is (and more recently Talk to the City) helped structure and process input across large numbers of participants, while in Libya the Digital Dialogues platform supported repeated large-scale, mobile-accessible deliberations with rapid analysis of participant input. Beyond these case studies, AI methods have also supported scaling in other ways, including in recruitment/sortition pipelines, agenda management, and information summarization.

While AI-based deliberation workflow tools promise to enable deliberation at greater scale, we risk centralizing control over key elements of the *substance* of deliberation, placing it in the hands of whoever designs the models, their inputs, and their outputs. Such risks can be exacerbated by *overreliance* on AI, a documented phenomenon in which people tend to defer by default to AI recommendations, even when these recommendations are suboptimal or unreliable. A key question, then, is how to use automation to expand the scale and frequency of deliberation while ensuring that power over the process is distributed intentionally and remains so even after accounting for how participants will change their behavior in response to the presence of these tools.

The case studies discussed earlier also illustrate how AI can contribute to ***sensemaking***: in Taiwan, Pol.is helped map agreement patterns and identify consensus in participants' responses; in the Habermas Machine, an LLM more explicitly synthesized across participants' inputs by iteratively generating and revising potential common-ground statements; and in Libya, an AI-assisted analysis pipeline organized open-text input into themes. Beyond these examples, there has recently been a rapid proliferation and deployment of AI models for performing sensemaking tasks.[71]

[71] Fish (2023); Kabbara (2025).

PROOFS

While sensemaking tools are highly promising, they are also embedded in the substance of deliberation and thus run the risk of becoming implicit architects of key choices in deliberation. For example, they can subtly steer attention, reframe issues, introduce biases, or collapse minority perspectives, all while appearing "objective" and "technically replicable." These risks pose another key challenge to future work: to empower participants with richer information and more incisive interpretations of others' ideas, while preserving their ability to inspect, challenge, and revise the interpretive choices that shape which ideas are elevated.

Across these cases, we also see AI being used to mitigate our third bottleneck, *unequal participation and inclusion*. In Libya, the Digital Dialogues were mobile-accessible and conducted in local dialects. In Taiwan, citizens could contribute online in a diversity of formats. These case studies suggest to a future in which multimodal interfaces and AI assistance lower barriers to participation and make participation less costly for citizens by fostering communication across language and geography. Artificial intelligence helps participants learn about issues more quickly and make inferences across diverse input formats, allowing people to contribute in the mode they find most natural.

When trying to elicit the input of a more diverse set of people at a lower participation cost, it is important to distinguish between techniques that make giving input easier, and those which use more sophisticated inference techniques to infer what people *would* say, without directly asking them. Proposals in the latter vein range from empowerment to *replacement*, with some proposals approaching the idea of replacing participants with AI-based proxies as a way of avoiding the time and effort costs of human participation. Aside from the question of whether it is technically possible for AI to act as a faithful human proxy, such proposals raise questions about of the very purpose of democratic participation: Is it simply a transfer of information or is it a personal act of investment in the public sphere? As our technological toolkit encompasses more tools aimed at more equal inclusion, one can question which elements of democratic participation a given tool empowers versus replaces – and how this may differ along the same (or different) societal dimensions on which inequality appears today.

Finally, these case studies offer a glimpse into the future of how AI can be used to enhance ***trust and perceived legitimacy***. In both Taiwan and Libya, AI tools helped produce outputs that were more legible to broader publics and more usable by decision-makers. In Libya, in particular, the dialogues were also broadcast on social media and live television, extending visibility beyond direct participants. Looking ahead, AI may help make citizen deliberation more accessible to the wider public by allowing members of the public to directly query detailed information about the process. On the other hand, the opacity and inherent technicality of AI models can shift trust from democratically accountable institutions and facilitators to inscrutable systems and their designers. When recommendations come from an AI-infused deliberation

PROOFS

pipeline, there is the risk that trust becomes more a matter of accepting the semi-automated pipeline that produced them, rather than considering the reasons of deliberation participants.

Taken together, these cases – and more broadly, the exploding landscape of deliberative technologies – illustrate why there is so much excitement about AI in this space: AI can offer toolkits that can help address some of the most stubborn bottlenecks in enmeshing more deliberative participation in our democratic institutions. Ultimately, however, whether AI deepens democracy will be decided not just by the technical sophistication of our tools but by the political choices that govern their use – whether they augment judgment or automate it; whether they make deliberation more legible or simply relocate trust from publics and institutions to technical systems; whether they reduce barriers to participation or reproduce inequality behind a more inclusive interface; and whether they strengthen public reasoning or bypass it in the name of efficiency.

1.6 CONCLUSION

The integration of generative AI in political campaigns, election administration, social movements, and citizen deliberation holds both great promise and the potential to damage the quality of our democracies. On one hand, AI potentially broadens and deepens democracy by engaging and empowering citizens, enriching the connections between citizens and their leaders, and facilitating sophisticated large-scale deliberation. These capabilities can empower resource-strapped campaigns, enhance the service quality of under-funded election offices, and provide marginalized social movements with new tools for mobilization and accountability. By addressing long-standing bottlenecks in participation and sensemaking, AI holds the promise of making democratic institutions more responsive and inclusive. However, these same tools can be weaponized to mislead voters, amplify illiberal politics, or centralize control in the hands of well-resourced platforms and technical elites.

No one can now say what the impact of AI on elections, campaigns, movements, and public deliberation will be. That future depends in part on the development of AI technologies and in part on how a vast array of political actors utilize these technologies to enhance their capabilities and achieve their ends. Will AI be open or closed? Biased and narrow in perspective or encompassing? Will the information these systems provide serve truth-seeking impulses or serve those who can pay the most or who seek domination? Will effective technologies amplify or flatten preexisting asymmetries of resources and power in the political sphere? The answers to these questions do not depend in the first instance on pure technological questions such as the sophistication of AI algorithms, the speed with which advanced chips can execute matrix multiplication, and the scale of data centers. Rather, the future of the democratic environment depends on regulatory frameworks like the EU AI Act, the competitive maneuvers of political and economic actors, and

PROOFS

the ways in which leaders integrate these tools into existing organizations and create new organizations. Ultimately, the impact of AI on the quality of democracy in the future depends upon how democracies decide to shape AI and its political uses now.

REFERENCES

ABP News Bureau. (2024, May 29). Alia Bhatt, Kareena Kapoor and other celebs support Gaza as "All Eyes on Rafah" trends after Israel's attack. *ABP News*. https://tinyurl.com/mrsbxens

Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, *33*(2), 3–30. https://doi.org/10.1257/jep.33.2.3

Allyn, B. (2024, June 3). "All Eyes on Rafah" is the internet's most viral AI image: Two artists are claiming credit. *New England Public Media*. https://tinyurl.com/4k4w9cwe

Al-Ubaydli, O., List, J. A., & Suskind, D. (2019). *The science of using science: Towards an understanding of the threats to scaling experiments* (NBER Working Paper No. 25848). National Bureau of *Economic Research*. https://doi.org/10.3386/w25848

Banerjee, D., Dutta, S., & Dutta Choudhury, S. (2024, June 5). *AI amplifies political reach but magnifies disinformation in India elections*. Asia Pacific Foundation of Canada. Accessed March 28, 2026. https://tinyurl.com/jj8z543f

Benford, R. D., & Snow, D. A. (2000). Framing processes and social movements: An overview and assessment. *Annual Review of Sociology*, *26*, 611–639. https://doi.org/10.1146/annurev.soc.26.1.611

Brennan Center for Justice. (2025a). *Local election officials survey – July 2025*. Accessed March 28, 2026. https://tinyurl.com/2tvbsz4j

Brennan Center for Justice. (2025b). *Local election officials survey – May 2024*. Accessed March 28, 2026. https://tinyurl.com/ymua2bhk

Burga, S. (2024, May 29). Why the "All Eyes on Rafah" AI post is going viral on social media. *Time*. https://time.com/6983344/all-eyes-on-rafah-ai-post-viral/

Burgess, A., & Morris, B. (2024, May 28). As the world reacts to Israel's deadly air strike on Rafah, here are key events from the Gaza war leading up to the attack. *ABC News*. https://tinyurl.com/3zbfkamp

Centre for Responsible Union AI. (n.d.-a). Beyond the hype: What union staff actually want from AI. Accessed January 18, 2026. www.agileunions.ai/p/what-union-staff-actually-want-from-ai

Centre for Responsible Union AI. (n.d.-b). How ABVV-Metaal used AI to automate creating and sending union leave letters for reps. Accessed March 28, 2026. www.agileunions.ai/p/automate-creating-sending-union-leave-letters

Centre for Responsible Union AI. (n.d.-c). How NASUWT used an AI chatbot to make their knowledge base more accessible to members. Accessed March 28, 2026. www.agileunions.ai/p/case-study-nasuwt-chatbot-members-resignation-notice-periods

Centre for Responsible Union AI. (n.d.-d). How PCS created an AI chatbot to support union reps practice conversations to recruit new members. Accessed March 28, 2026. https://tinyurl.com/bdcsjcem

Cohen, J., & Kugelberg, H. D. (2025, May 5). Trust in AI mediators may change deliberative outcomes [eLetter]. *Science*. https://philpapers.org/archive/COHTIA-2.pdf

PROOFS

Cortés, E., Norden, L., Frase, H., & Hoffmann, M. (2023). *Safeguards for using artificial intelligence in election administration*. Brennan Center for Justice. Accessed April 11, 2026. https://tinyurl.com/3ynjj696

Cushing Rodriguez, S. (2023, October 4). Consensus building in Taiwan, the poster child of digital democracy. *Democracy Technologies*. Accessed March 28, 2026. https://democracy-technologies.org/participation/consensus-building-in-taiwan/

Dahl, R. A. (1990). *After the revolution? Authority in a good society*. Yale University Press.

Dunworth, L. (2024, May 29). Dua Lipa shares support for #AllEyesOnRafah: "Burning children alive can never be justified." *NME*. www.nme.com/news/music/dua-lipa-shares-support-for-all-eyes-on-rafah-3760399

Elliott, V. (2024, June 11). There's an AI candidate running for parliament in the UK. *Wired*. Accessed March 28, 2026. www.wired.com/story/ai-candidate-running-for-parliament-uk/

Fish, S., Gölz, P., Parkes, D. C., Procaccia, A. D., Rusak, G., Shapira, I., & Wüthrich, M. (2023). *Generative social choice* (arXiv:2309.01291) [Preprint]. arXiv. https://arxiv.org/abs/2309.01291

Flanigan, B., Gölz, P., Gupta, A., Hennig, B., & Procaccia, A. D. (2021). Fair algorithms for selecting citizens' assemblies. *Nature*, *596*(7873), 548–552. https://doi.org/10.1038/s41586-021-03788-6

Foos, F. (2024). The use of AI by election campaigns. *LSE Public Policy Review*, *3*(3), 4–5. https://doi.org/10.31389/lseppr.112

Foos, F., John, P., Valiente-Jorda, F., & Cinar, A. C. (2026). *Scaling-up challenges in political interventions* [working paper].

For this Hillary Clinton listening tour, few hugs but much note-taking. (2015). *The New York Times*. Accessed March 28, 2026. https://tinyurl.com/mwh8dxh3

Freed, B. (2020, September 17). Election questions? Idaho's new Watson-powered chatbot has answers. *StateScoop*. Accessed March 28, 2026. https://statescoop.com/election-chatbot-idaho-watson-ai/

Fung, A., & Lessig, L. (2023, June 2). How AI could take over elections – and undermine democracy. *The Conversation*. https://tinyurl.com/yzcfjm2h

Germann, M., Marien, S., & Muradova, L. (2024). Scaling up? Unpacking the effect of deliberative mini-publics on legitimacy perceptions. *Political Studies*, *72*(2), 677–700. https://doi.org/10.1177/00323217221137444

Green, D. P., & Gerber, A. S. (2019). *Get out the vote: How to increase voter turnout* (4th ed.). Brookings Institution Press.

Greschler, G. (2024, February 27). Voters are being called by campaigns' AI bots. I talked to one for 10 minutes. *San Francisco Standard*. https://sfstandard.com/2024/02/27/peter-dixon-congress-ai-robocalls-civox/

Guevara, F. (2025, March 4). Using AI now to improve movements' effectiveness: A basic introduction for social activists [Commentary]. Ash Center for Democratic Governance and Innovation. https://tinyurl.com/y2hczp73

Guglielmo, C. (2023, October 21). AI and you: NYC mayor can't really speak Mandarin, the AI money trail, who sets the rules. *CNET*. Accessed March 28, 2026. https://tinyurl.com/26vw8zpj

Hackenburg, K., Tappin, B. M., Hewitt, L., et al. (2025). The levers of political persuasion with conversational artificial intelligence. *Science*, *390*(6777), Article eaea3884. https://doi.org/10.1126/science.aea3884

PROOFS

Hall, A. (2026, March 26). Building political superintelligence. *Substack*. https://substack.com/home/post/p-192214946

Imran Khan uses A.I. to give victory speech in Pakistan. (2024, February 11). *The New York Times*. Accessed March 28, 2026. https://tinyurl.com/4pjzrenu

Irwin, C., Masood, D., Wählisch, M., & Konya, A. (2021). *Using artificial intelligence in peacemaking: The Libya experience*. WAPOR 74th Annual Conference (virtual conference). Accessed July 7, 2026. https://peacepolls.etinu.net/peacepolls/documents/009260.pdf

Issenberg, S. (2013). *The victory lab: The secret science of winning campaigns*. Crown.

Jie, J. (2025, March). Unlocking AI's potential for social movements. Ash Center for Democratic Governance and Innovation. https://ash.harvard.edu/articles/unlocking-ais-potential-for-social-movement/

Jungherr, A., Rauchfleisch, A., & Wuttke, A. (2026). Artificial intelligence in election campaigns: Perceptions, penalties, and implications. *Political Communication*, 42(4), 585–606. www.tandfonline.com/doi/full/10.1080/10584609.2025.2611913

Kabbara, J., Phan, T. M., Rakhilin, M., Detwiller, M. E., Dimitrakopoulou, D., & Roy, D. (2025). AI-assisted sensemaking: Human-AI collaboration for the analysis and interpretation of recorded facilitated conversations. In *Extended abstracts of the CHI conference on human factors in computing systems* (pp. 1–8). Association for Computing Machinery. https://dl.acm.org/doi/full/10.1145/3706599.3706688

Kalla, J. L., & Broockman, D. E. (2016). Campaign contributions facilitate access to congressional officials: A randomized field experiment. *American Journal of Political Science*, *60*(3), 545–558. https://doi.org/10.1111/ajps.12180

Lin, H., Czarnek, G., Lewis, B., et al. (2025). Persuading voters using human–artificial intelligence dialogues. *Nature*, *648*(8093), 394–401. https://doi.org/10.1038/s41586-025-09771-9

List, J. A. (2022). *The voltage effect: How to make good ideas great and great ideas scale*. New York: Crown Current, a division of Random House.

Mann, C. B. (2021). Can conversing with a computer increase turnout? Mobilization using chatbot communication. *Journal of Experimental Political Science*, *8*(1), 51–62. https://doi.org/10.1017/XPS.2020.5

National Conference of State Legislatures. (2026, March 24). Artificial intelligence in election administration. Accessed March 29, 2026. https://tinyurl.com/4skrs9yf

Neblo, M. A., Esterling, K. M., & Lazer, D. M. J. (2018). *Politics with the people: Building a directly representative democracy*. Cambridge University Press.

New York State Senate. (2023). Senate Bill S9678, 2023–2024 legislative session: Relates to materially deceptive media in political communications. Accessed March 29, 2026. www.nysenate.gov/legislation/bills/2023/S9678/amendment/B

Offer-Westort, M., et al. (2026). Deep canvassing with automated conversational agents: Personalized messaging to change attitudes. *Research & Politics*, *13*(1), Article 20531680251414926. https://doi.org/10.1177/20531680251414926

OpenAI bans developer of Dean Phillips bot. (2024, January 20). *The Washington Post*. www.washingtonpost.com/technology/2024/01/20/openai-dean-phillips-ban-chatgpt/

Orrick. (n.d.). U.S. AI law tracker. https://ai-law-center.orrick.com/us-ai-law-tracker-see-all-states/

Participedia. (n.d.-a). 21st-century town meeting. Accessed March 26, 2026. https://participedia.net/method/145

PROOFS

Participedia. (n.d.-b). AmericaSpeaks. Accessed March 26, 2026. https://participedia.net/organization/americaspeaks

Partnership for Public Service. (n.d.). Leadership program inspires an AI revolution at the State Department. Accessed March 28, 2026. https://tinyurl.com/mvpydavv

PBS News. (2024, February 27). Chatbot info on U.S. elections is inaccurate, misleading and could keep voters from polls, report finds. *PBS News*. Accessed March 28, 2026. https://tinyurl.com/bdhryuut

Peterson, T. (2025, November 4). How The Times is using AI to model synthetic focus groups from human audiences. *Digiday*. Accessed March 28, 2026. https://tinyurl.com/ye5pfhpp

Picón, I. (2025, March 6). AI-powered bots and electoral participation: A view from the Venezuelan experience. Ash Center for Democratic Governance and Innovation. https://tinyurl.com/3y779w6y

Pow, J., Van Dijk, L., & Marien, S. (2020). It's not just the taking part that counts: "Like me" perceptions connect the wider public to mini-publics. *Journal of Deliberative Democracy*, *16*(2), 43–55. https://doi.org/10.16997/jdd.395

Reuters. (2016, May 28). France's Macron launches door-to-door campaign as presidency battle looms. *Reuters*. Accessed March 28, 2026. https://tinyurl.com/47us44nw

Sakunia, S. (2024, July 12). AI and deepfakes played a big role in India's elections. *New Lines Magazine*. Accessed March 28, 2026. https://tinyurl.com/ystxw4r7

Samuels, B. (2023, August 14). How a startup is using AI to write fundraising emails. *The Hill*. Accessed March 28, 2026. https://tinyurl.com/4eb6h6ep

Schmidt, S., Rich, S., Herrero, A. V., & Paúl, M. L. (2024, August 4). Maduro lost election, tallies collected by Venezuela's opposition show. *The Washington Post*. www.washingtonpost.com/world/2024/08/04/maduro-gonzalez-election-actas-analysis/

Schneier, B., & Sanders, N. E. (2025). *Rewiring democracy: How AI will transform our politics, government, and citizenship*. The MIT Press.

Schneier, B., & Sanders, N. E. (2026, January 11). Rewiring democracy now: A new kind of political engagement emerges in Japan. *The Renovator*. https://therenovator.substack.com/p/rewiring-democracy-now

Silicon Valley insiders are trying to unseat Biden with help from AI. (2024, January 18). *The Washington Post*. Accessed March 28, 2026. www.washingtonpost.com/elections/2024/01/18/ai-tech-biden/

Stewart, C., III. (2022). Trust in elections. *Daedalus*, *151*(4), 234–253. https://doi.org/10.1162/daed_a_01953

Tessler, M. H., Bakker, M. A., Jarrett, D., (2024, October 18). AI can help humans find common ground in democratic deliberation. *Science*, *386*(6719), Article eadq2852. https://doi.org/10.1126/science.adq2852

Trump backers inundate election offices with requests for 2020 records. (2022, September 11). *The Washington Post*. Accessed March 28, 2026. www.washingtonpost.com/nation/2022/09/11/trump-election-deniers-voting/

Tseng, Z. (n.d.). Amplifying voices: Talk to the City in Taiwan. Objectives Institute. Accessed March 28, 2026. https://ai.objectives.institute/blog/amplifying-voices-talk-to-the-city-in-taiwan

UK Parliament. (n.d.). Election result for Brighton Pavilion (constituency). Accessed March 28, 2026. https://members.parliament.uk/constituency/3945/election/422

U.S. Election Assistance Commission. (n.d.). *Signature verification and cure process*. Accessed on April 11, 2026. https://tinyurl.com/5u5cjp2n

PROOF

Ulrich, S. (2025, February 3). Using data to expose systemic gender-based violence – in 10 steps. *Global Investigative Journalism Network*. https://gijn.org/stories/10-steps-data-expose-gender-violence-eswatini/

Westwood, S., Grimmer, J., & Hall, A. (2025, May 8). *Measuring perceived slant in large language models through user evaluations* [Working paper]. Stanford Graduate School of Business. https://tinyurl.com/yc394dhn

Yoder, K., & Tan, A. (2024, June). *CEIR focus brief: Election officials & the misuse of public records requests*. The Center for Election Innovation & Research. Accessed March 28, 2026. https://tinyurl.com/4z5h6ex9

PROOFS

2

Easy to Produce, Hard to Persuade

The Asymmetric Effects of AI on the Online Information Ecosystem

Brendan Nyhan, Jennifer Pan, Alexandra Siegel, and Yamil Velez

PROOFS

The rapid increase in the usage and capacity of artificial intelligence (AI) has raised widespread concerns about information manipulation. Leading figures worry that large language models (LLMs) like ChatGPT will make it possible to create seemingly authentic false or misleading content that is tailored for personalized persuasion and that exposure to this content will shape public beliefs and attitudes at scale.

For instance, in US Senate testimony, OpenAI CEO Sam Altman said, "It's one of my areas of greatest concern: the more general ability of these models to manipulate, to persuade, to provide sort of one-on-one, you know, interactive disinformation" (Altman 2023). Anthropic CEO Dario Amodei similarly worried in a recent essay that "[m]uch more powerful versions of these models … would likely be capable of essentially brainwashing many (most?) people into any desired ideology or attitude" (Amodei 2026). Political figures have expressed similar concerns. US Senator Richard Blumenthal asserted before the 2024 election that "a deluge of deception, disinformation, and deepfakes are about to descend on the American public" via "political ads and other forms of disinformation that are made possible by artificial intelligence" (Blumenthal 2024). Former South Korean President Yoon Suk Yeol warned that "[f]ake news and disinformation based on artificial intelligence and digital technology not only [violate] individual freedom and human rights but also [threaten] democratic systems" (Park and Kim 2024). And United Nations Secretary-General António Guterres told the Security Council that "[t]he advent of generative AI could be a defining moment for disinformation and hate speech – undermining truth, facts and safety, adding a new dimension

We thank Nate Persily, Joshua Tucker, and Rachel Gillum for helpful comments. AI tools were used to find and cite sources, to provide feedback on draft versions of the manuscript, and for suggestions on the title. All errors are our own.

Tucker_9781047774734_C002.indd 53 15-09-2026 07:03:45

to the manipulation of human behaviour and contributing to polarization and instability on a vast scale" (Guterres 2023).

Governments, technology platforms, and civil society should, of course, guard against potential threats from the misuse of AI, but we urge caution among those rushing to conclusions about its effects on the public's beliefs and attitudes. The fears being expressed about the effects of AI mirror those offered after the 2016 US presidential election about the harmful political effects of social media and other forms of online information exposure. Many of the claims made at the time did not hold up to scrutiny. After almost a decade of research, scholars have found that algorithms do not appear to cause mass polarization (e.g., Guess et al. 2023); that most people in the US and other Western countries are not in so-called echo chambers dominated by pro-attitudinal sources (e.g., Guess 2021; Arguedas et al. 2022); and that relatively few people consume most of the false or extremist content seen online (Budak et al. 2024). Similarly, despite extensive warnings about the potential role of deepfakes and other forms of AI-based misinformation in the 2024 US election, they played little role in the campaign (Schneier and Sanders 2024). Instead, we have seen the proliferation of "AI slop" that is readily recognized as fake and instead seeks to appeal to the emotions of partisan audiences (Jingnan 2024; Bond and Brumfiel 2025).

Importantly, however, these conclusions may differ outside of the West, where less is known about the prevalence and effects of online misinformation (Blair et al. 2024; Budak et al. 2024). The harms from AI may be greater, for instance, in countries with lower media and technological literacy or higher levels of ethnic tension. We are most concerned about the potential harms that AI tools can create by relaxing monitoring, production, and other capacity constraints that previously limited the reach and scope of information influence and control efforts. First, AI lowers barriers for state and nonstate actors to conduct information campaigns that were previously too costly to undertake. Second, in authoritarian countries like China, LLM companies operating under de facto state control are developing models that reflect the regime's political objectives, risking the embedding and spread of these perspectives as emergent properties of the technology.

In this chapter, we examine how AI affects exposure to information and persuasion, including on social media. We begin at the individual level by examining how exposure to political information may be affected by AI platforms and content, as well as the potential persuasive effects of AI content on public beliefs and attitudes. We next consider how AI may change the incentives and activities of technology platforms as well as state and nonstate actors who seek to influence the political and social world. Finally, we describe how LLMs and other forms of AI enable new research designs that can answer fundamental questions about information exposure and persuasion.

PROOFS

2.1 INFORMATION EXPOSURE

As generative AI use expands rapidly around the world, particularly in middle- and low-income countries, it is reshaping how people are exposed to digital information (Chatterji et al. 2025). However, changes in information exposure matter for political processes and outcomes only to the extent that they facilitate attitudinal or behavioral change among those exposed to it. We expect AI to accelerate existing patterns in these respects, rather than transform them. Most people will continue to consume relatively little political content, regardless of how it is delivered. Similarly, the challenges of persuasion at scale persist whether content is curated by humans, traditional machine-learning algorithms, or AI-enhanced systems.

First, AI is changing patterns of information exposure when it comes to search. Specifically, online search is shifting from a process of searching with keywords to find relevant sources to interacting with an AI intermediary that provides summaries and serves as a gatekeeper to original sources. For example, AI-generated summaries on search engines like Google have dramatically increased "zero-click searches" – when users obtain answers without clicking through to an external website. Between May 2024 and May 2025, the percentage of zero-click searches jumped from 56 percent to 69 percent (Thekkethil 2025). Organic traffic to news and media websites dropped by 600 million visits from 2024 to 2025 (Similarweb 2025). In addition, an increasing number of people are turning directly to AI platforms to get answers via chat interfaces. As of July 2025, approximately 10 percent of the global adult population (700 million users) was using ChatGPT weekly (Chatterji et al. 2025). Compared to traditional search, AI search draws on a narrower range of sources, produces less varied responses, and disproportionately surfaces lower-credibility sources (Aral, Li, and Zuo 2026).

On social media, the integration of generative AI expands the set of behavioral signals used to inform platform operations. Social media platforms have long used machine learning with behavioral signals, such as likes, shares, watch time, scroll speed, and follow patterns to support operational decisions ranging from content moderation and spam detection to ranking and recommendation. Even prior to generative AI, platforms followed TikTok in shifting from follower graph-based feeds to algorithmically curated feeds, which expanded the scope of potential exposure by surfacing content outside users' social networks (Sanderson, Messing, and Tucker 2024). The integration of generative AI creates new sources of behavioral data, such as users' conversations with platform-embedded AI chatbots, and allows for richer analysis of content. For example, Meta announced in October 2025 that it would use data from interactions with AI chatbots across Facebook, Instagram, and WhatsApp to personalize content and ads targeting the user (Meta 2025). TikTok uses multi-modal LLMs to extract specific claims from videos to curate content, and

PROOFS

Instagram indexes video, text, and audio content to improve reach (TikTok 2024; Wilson 2026). However, these changes are largely incremental. There is little reason to expect them to overcome fundamental barriers – such as attention paid to politics and the persuasiveness of content – that would lead to attitudinal and behavioral shifts at scale.

Beyond these more direct impacts on information exposure, generative AI may also affect downstream information exposure by shifting patterns of information production. Artificial intelligence tools lower creation costs, potentially enabling more content to be produced with less effort by humans. Such a change could increase the productivity of journalistic organizations, though many observers worry about inaccuracy and degradation of the quality of news reporting (Setty 2026). In addition, autonomous AI agents can now contribute to open-source code, publish content online, and communicate with other AI agents as well as humans without direct human oversight. In February 2026, an AI agent wrote and disseminated a personalized attack piece on Scott Shambaugh, a matplotlib developer, by researching Shambaugh's commit history on GitHub after Shambaugh rejected its code contribution (Shambaugh 2026). Platforms like Moltbook, a social network exclusively for AI agents, which claims over 1.4 million registered agents as of early 2026 (Yildiz 2026), represent an emerging space where AI-to-AI interaction generates its own information ecosystem largely outside human oversight.

The growing volume of AI-generated information can displace human knowledge creation. Within six months of ChatGPT's release, activity on Stack Overflow decreased 25 percent in countries where ChatGPT was accessible compared to countries where access to ChatGPT was limited (del Rio-Chanona et al. 2024). A similar pattern was observed on Stack Exchange, where questions dropped by 18 percent (Quinn and Gutt 2025). This decrease may reflect efficiency gains, as users get answers immediately without needing to wait for community responses, but these shifts mean the incentive to create or contribute original human-generated material declines, potentially diminishing the publicly available base of human-generated information, even as the volume of content in circulation is greatly boosted by generative AI.

Such changes in the information environment will not necessarily have any particular political effects. However, as information volume grows, people will also increasingly rely on AI systems to curate information for them, making LLMs a gatekeeper as well as a producer. When users turn to ChatGPT or other models instead of social media or news sites, LLMs decide what is relevant, what to summarize, and what to omit from an information environment that these models are increasingly populating.

Importantly, though, these dynamics may have different effects across linguistic contexts. Generative AI systems perform worse for low-resource languages (e.g., Burmese, Armenian, Yoruba) than for high-resource languages (e.g., English, French, Chinese) across a range of tasks, including translation, factual question answering, text classification, and summarization,

PROOFS

primarily due to the lack of training data (Robinson et al. 2023; Li et al. 2025). Multilingual LLMs display information biases when queried in low-resource languages and, instead, draw on high-resource language sources which reinforce those linguistic perspectives and marginalize low-resource language viewpoints (Sharma, Murray, and Xiao 2025). In addition, LLMs are more likely to produce harmful or unsafe output in low-resource languages due to less robust safety training and alignment (Yong et al. 2023; Deng et al. 2024). While newer models have expanded their language coverage, most major LLMs underperform for low-resource languages (Pava et al. 2025). The state of language resources depends on historical patterns related to institutions and development. Languages of former colonial powers and wealthier countries have far more digital text available for AI training than languages spoken primarily in lower-income countries or by marginalized populations. For example, Hindi is not a high-resource language despite the large population of speakers because English dominates India's digital sphere. In addition, the digital data that LLMs are trained on reflect the information environment of source countries. Research shows that text scripted for the Chinese media by the Chinese Communist Party appears in LLM training datasets and that prompting models that were developed outside of China in Chinese generates more positive assessments of China's political institutions than the same prompts in English (Waight et al. 2026). The global shift to AI-influenced information exposure will reinforce these existing inequalities. The quality and relevance of information may be systematically lower for populations who already face political and economic barriers to reliable information, while those who can produce digital content at scale may see their influence strengthened. These disparities are rooted in institutions and socioeconomic structures. Understanding and addressing these dynamics requires the engagement of social science that purely technical approaches risk overlooking.

2.2 PERSUASION AND BELIEF/ATTITUDE CHANGE

New technologies ranging from radio to television to video games have frequently inspired panic around their potential harms, including fears of brainwashing and propaganda (e.g., Starker 1989; Drotner 1999; Orben 2020). In this sense, the fears being expressed about the persuasive powers of AI are familiar. As with those technologies, AI-generated content can indeed be persuasive and lead to durable attitude changes, but these effects are likely to be smaller than people fear and to vary with exposure, as we discuss in this section.

We first distinguish between AI source effects and AI content effects. In general, when identified as the source, AI has mixed or inconsistent effects that may vary by domain or other contextual factors. Lu, Tormala, and Duhachek (2025) find people are more open to counter-attitudinal persuasion from AI rather than from human sources because AI is seen as more informative and less biased. By contrast, Boissin et al. (2025), Gallegos et al. (2025), and Spearing

PROOFS

et al. (2025) find no effect of AI versus human attribution in persuasive messages (including misinformation). And Altay and Gilardi (2024), Nanz et al. (2025), and Wittenberg et al. (2025) find people are more skeptical of journalistic content, media outlets, and images labeled as AI-generated, respectively.

Some studies examine political slant and bias in AI content, instead of, or in addition to, testing the causal effects of exposure to it. Though definitional and measurement concerns persist (Fisher et al. 2025a), most studies find content from major LLMs leans left, with the possible exception of xAI's Grok, which has shifted to the right (e.g., Rozado 2024; Motoki et al. 2025; Rozado 2025; Westwood, Grimmer, and Hall 2025; Wirtschafter and Nadgir 2025). Exposure to LLM content may, accordingly, have some effect on people's political attitudes. Potter et al. (2024) find, for instance, that engaging in a conversation with an LLM about former US President Joe Biden and current President Donald Trump shifted people's views toward Biden. However, absent a neutral or non-AI baseline, experimental designs like these do not isolate the causal effect of slant in LLM content or the information provided about the candidates.

The effects of AI-generated content are most precisely estimated in studies that manipulate exposure to AI content on a topic of interest. These studies often find that exposure to AI content or participation in AI dialogues can generate durable changes in people's political beliefs and attitudes – changes that often match or outperform those created by exposure to human-created static content (e.g., Costello et al. 2024; Hackenburg and Margetts 2024; Bai et al. 2025; Costello et al. 2025a; Costello et al. 2025b; Czarnek et al. 2025; Lin et al. 2025; Fisher et al. 2025b). However, these findings differ in magnitude from the claims being made about the persuasive powers of AI. For instance, the most high-profile study in this area is Costello et al. (2024) which finds that belief in the conspiracy theory debunked by an AI in a dialogue with online survey participants was reduced by approximately 14 percentage points (and other studies like Summerfield et al. 2025, have found smaller effects, averaging around 7 percentage points).

Findings differ on the factors that make AI content or dialogue exposure persuasive. One theory is that LLMs can tailor the content they produce to attributes of the person receiving the information. Two recent studies find that personalized content from LLMs is more persuasive than non-personalized content (Matz et al. 2024; Salvi et al. 2025). However, two other studies find few differences resulting from personalization (Hackenburg and Margetts 2024; Argyle et al. 2025), and a third finds the effect is small relative to other factors (Hackenburg et al. 2025).

An alternative account emphasizes the role of novel facts (or fact-like claims) in AI persuasion. Notably, Costello et al. (2025c) uses an ablation design to probe the mechanism for the belief change findings in Costello et al. (2024), which shows how conversations with AI can reduce conspiracy belief. Of the potential mechanisms tested, the only one that eliminated the observed effect

PROOFS

was an experimental condition in which the LLM was told not to provide rational, evidence-based arguments. Persuasive effects were also much smaller in a no-facts condition evaluated by Bai et al. (2025). The use of facts and evidence by LLMs was, in turn, associated with greater participant belief change in Bai et al. (2025), Costello et al. (2025a), and Lin et al. (2025). The facts provided by the LLMs in studies like these have been found to be largely accurate when evaluated (Costello et al. 2024; Lin et al. 2025). In some cases, such facts may be lacking; Costello et al. (2025c) finds that the LLM instead urged caution before facts were known in challenging conspiracy theories about the 2024 assassination attempts against Donald Trump. However, if pushed further, models may make dubious claims that are persuasive but not accurate. Hackenburg et al. (2025) shows that post-training of LLMs that is intended to increase persuasion tends to degrade accuracy, suggesting a key tradeoff in which demands for more factual or fact-like information to support a claim cause models to veer into inaccuracy.

Importantly, the key role of facts and evidence in AI persuasion does not mean that AIs cannot misuse information or mislead people. Model accuracy has tended to improve over time, but such increases are not necessarily linear in time. Notably, Hackenburg et al. (2025) found that GPT-4.5 was less accurate than prior GPT models tested despite being much larger and having more post-training. Accuracy may also degrade in certain contexts. For instance, Lin et al. (2025) found that LLMs made more inaccurate claims when prompted to make an argument for conservative candidates.

In particular, it has long been clear that LLMs can be used to create propaganda, misinformation, and other forms of malign persuasive content (Goldstein et al. 2024). Indeed, they already appear to be in use to expand the output of Russian-backed state media (Wack et al. 2025). The potential for misuse is substantial. Hackenburg et al. (2025) shows that post-training intended to increase persuasion is feasible and can make a smaller open weights model (Llama-8B) as persuasive as to a larger proprietary model (GPT-4.0). Costello et al. (2026) similarly shows that LLMs are as effective in convincing people to believe in conspiracy theories as they are in debunking them. Measuring and mitigating these risks is an important topic for future research (Kowal et al. 2025).

Countermeasures such as inoculation may not be enough to undo the effects of exposure to AI-generated misinformation (Spearing et al. 2025). First, accurate identification of such content at scale is difficult even for platforms. People may be even worse at making such distinctions, including for images (DeVerna et al. 2024). In addition, reminders of the prevalence of AI images can increase general skepticism about news and digital media (Campante et al. 2025; Sanderson, Zhong, and Tucker 2025).

It is also important to note that parasocial relationships with AI companions could expand the scope of AI persuasion by deepening attention and trust, which often constrain persuasive effects at scale. Contemporary LLM chatbots

PROOFS

are responsive, remember information, and can sustain long, emotionally salient conversations, which can increase disclosure, perceived intimacy, and reliance over time (Smith et al. 2025). This sustained engagement can create opportunities for iterated persuasion – testing messages, adapting to objections, and calibrating framing to the user's goals. As a result, even if average "one-shot" exposure effects are small, influence effects may compound over time among high engagement users who treat the chatbot as a trusted social partner (Qi et al. 2025).

We conclude by noting two challenges for research in this area. The first is that the temporal validity of these findings may be low due to rapid change in the state of the art in LLMs and other forms of generative AI (Munger 2019; Munger 2023). Human performance at detecting AI-generated images, for instance, is likely to degrade as technology improves. Second, we may observe second-order effects that alter the dynamics of persuasion in this domain. People may come to distrust AI content more or adjust their information diets to rely more on trusted sources (Campante et al. 2025). Artificial intelligence may also create or contribute to a so-called liar's dividend in which people use AI to deny the validity of true information (Sanderson, Messing, and Tucker 2024; Schiff et al. 2025). However, so far, people still do not instinctively disbelieve the images and videos they encounter as a result of AI awareness. Videos and images from Minneapolis, Minnesota, were widely circulated during the period that a large number of Immigration and Customs Enforcement (ICE) and Customs and Border Protection agents were deployed there and treated as credible (though, of course, AI-manipulated images and videos were also posted online). The federal government did not attempt to claim that authentic videos and images were AI creations. One constraint may be the number of videos that were created of the same incidents. In addition, news organizations are increasingly devoting resources to verifying digital media and certifying it as legitimate (e.g., Lum and Willis 2026).

2.3 PLATFORMS, STATE ACTORS, AND NONSTATE ACTORS

Social media reshaped political information environments largely by lowering distribution costs and barriers to entry, making it cheap and easy to reach large audiences, coordinate supporters, and exploit platforms for amplification (Zhuravskaya, Petrova, and Enikolopov 2020; Papathanassopoulos and Giannouli 2025). By contrast, generative AI lowers the costs of production and interpretation (Sanderson, Messing, and Tucker 2024). Creating tailored text, images, and narratives and classifying and distilling meaning from large quantities of content is faster and less costly than ever. These features of generative AI can be exploited by state and nonstate actors to manufacture, adapt, and summarize information at scale (Marcellino et al. 2023). In this section, we describe how generative AI is changing how state and nonstate actors can influence political information.

PROOFS

First, LLMs facilitate continuous production and adaptation of targeted information. In the social media era, even sophisticated influence operations often struggled to generate enough plausible, locally tailored content across languages and personas to reach their strategic goals. Generative AI reduces that constraint by enabling high-volume, stylistically diverse, rapidly iterated messaging, which is especially valuable when actors must constantly adapt to moderation and shifting narratives (Marcellino et al. 2023). Analyses of a recent Russian state influence campaign suggest that the adoption of generative AI coincided with expanded output while maintaining apparent persuasiveness, consistent with the emergence of more "industrialized" production pipelines (Wack et al. 2025). Similarly, experiments show that current models can generate high-quality components of election disinformation operations, lowering the time and skill requirements for running campaigns that previously demanded larger teams and specialized expertise (Williams et al. 2025). These changes benefit both state and nonstate actors but especially smaller, lower-resourced groups who can now run continuous real-time operations with limited human staffing (Pauwels 2024).

Second, state control over LLM development risks embedding regime narratives into the models themselves. This differs from social media–era information control, which operated primarily through censorship and propaganda directed at content after its creation. Large language model companies operating under de facto state control develop models that reflect the regime's political objectives from the outset. Studies of Chinese, Russian, and US-based LLMs show how this manifests as both hard refusals and soft omission or steering of tailored content to domestic audiences and languages (Noels et al. 2025; Pan and Xu 2026), with more fine-grained audits documenting semantic suppression in responses to sensitive prompts (Qiu, Zhou, and Ferrara 2025).

Importantly, state influence on model behavior is not limited to authoritarian regimes. Democratic governments are increasingly using procurement, regulation, and political pressure rather than direct ownership or formal party-state control to shape model development. In the US, federal acquisition guidance directs agencies to manage risks in procured AI systems and requires vendor transparency and testing/disclosure practices, which can incentivize providers to modify model behavior and safety features to meet contracting requirements. In 2025, the executive branch explicitly tied federal procurement to contested "ideological bias" standards to prevent "woke AI in the federal government" (Trump 2025). The next year, the Trump administration went further, designating Anthropic as a "supply chain risk" (Lawler and Curi 2026).

Third, the data used to train LLMs create the opportunity for political influence beyond national borders. Authoritarian regimes influence their domestic information environments, which then shape the training data used by models developed both inside and outside authoritarian settings (Yang and Roberts 2023). Waight et al. (2026) demonstrate that propaganda from political institutions already influences the output of US-based large language models via

PROOFS

their training data. Regime narratives can thus spread as emergent properties of the technology – not just through direct manipulation of foreign models but also via upstream contamination of the data ecosystem on which all models depend. State actors have also attempted more direct interventions, including industrial-scale content flooding designed to contaminate the web and the materials that retrieval-augmented generation (RAG) systems rely on, such as campaigns associated with pro-Kremlin ecosystems like the "Pravda" network (Danet 2025; Châtelet 2025). Some work cautions that apparent "poisoning" may instead reflect data voids (Golebiewski and Boyd 2018) – when high-quality information is scarce, low-quality or partisan sources can dominate what retrieval systems surface without requiring coordinated manipulation (Alyukov et al. 2025). These dynamics could benefit nonstate actors as well; even if small groups cannot poison closed training data pipelines, they may flood the open web or exploit data voids in ways that influence what systems retrieve and cite (Alyukov et al. 2025; Pauwels 2024).

Fourth, LLMs make personalization cheap and interactive in a way social media rarely did at scale. Social media enabled microtargeting and experimentation, but sustained, conversational persuasion or recruitment typically required humans in the loop (call centers, troll farms, volunteer brigades) or relatively rigid scripts. Large language models support interactive messaging that can respond to objections, maintain persona consistency, and tailor appeals across channels (Pauwels 2024). In this way, nonstate actors and extremist networks using language models can emulate interactive radicalization content and generate persuasive text cheaply, raising recruitment and mobilization risks even when platforms attempt enforcement (McGuffie and Newhouse 2020; Puczyńska et al. 2024). These changes also affect state actors because the same interactivity can be used to frame narratives and to scale harassment and intimidation efforts both domestically and transnationally (Sen and Farooq 2025). However, attention and trust still constrain the effects of these efforts. As a result, they may not result in opinion change even as they create expanded capacity for persistent engagement, testing, and targeting (Williams et al. 2025; Pauwels 2024; Sen and Farooq 2025). While limited attention may constrain average opinion change, it does not similarly constrain the effectiveness of harassment. Even when targets do not believe the message, persistent, personalized abuse can impose costs, such as time, stress, and safety concerns, that deter activism and participation. While coordinated trolling has long been a common occurrence on social media (Bradshaw and Howard 2017), LLMs reduce the labor required to sustain high-volume, adaptive harassment – including rapid replies, persona consistency, and cross-platform variation – potentially increasing prevalence and persistence. As with coordinated trolling operations, such harassment campaigns can be used to target a wide range of individuals from politicians to activists and everyday citizens.

Fifth, generative AI changes detection and attribution, allowing for greater plausible deniability by both state and nonstate actors. Social media

PROOFS

manipulation often leaves traces of coordination in metadata that platforms and researchers could detect. Large language models can generate endless paraphrasing and stylistic variants, making "copy-pasta" detection more difficult and raising the verification burden for defenders (Marcellino et al. 2023). These challenges interact with manipulation at the training data level. Even if individual posts are removed, the broader information environment may still be saturated, and retrieval systems may keep rediscovering variants (Danet 2025; Alyukov et al. 2025). As a result, those seeking to combat influence operations in the age of generative AI must focus not only on removing content but on maintaining the integrity of the evidence and reference ecosystems that models consult.

Finally, generative AI may intensify the cat-and-mouse nature of authoritarian and antiauthoritarian uses. Social media already empowered citizen journalism and mobilization while exposing activists to surveillance, repression, harassment, and doxxing. Generative AI perhaps intensifies this dynamic by improving surveillance, repression, and propaganda, as well as verification and documentation. On the one hand, AI tools can amplify state surveillance capacity by lowering the costs of monitoring vast quantities of social media posts, messages, images, and geolocation traces so that security services can identify networks and flag individuals quickly (CEIP 2019). Artificial intelligence–enabled biometric identification (especially facial recognition) can also facilitate the tracking of protesters and dissidents, increasing arrest and retaliation risks even when participation is brief or decentralized (Freedom House 2023). Importantly, these tactics are not limited to authoritarian contexts and have been increasingly utilized by democratic regimes. For example, recent reporting demonstrates that US ICE is using AI-enabled tools to analyze driver's license scans; extract information from seized phones; cross-search location data; and link records from federal, state, and commercial databases (Hubbard 2025). Similarly, local US law enforcement including the Miami-Dade Sheriff's Office and the Los Angeles Police Department have bought access to GeoSpy, an AI tool that can almost instantly geolocate where a photo was taken using features in the image, such as architecture and vegetation, and plan to use it in criminal investigations (Cox 2026).

On the other hand, AI can also make human rights work more agile by automating repetitive tasks, such as translation, summarization, and pattern identification, and by supporting digital tools for advocacy and monitoring (Espiritusanto et al. 2024). By lowering the costs of producing public facing content, generative AI can also help small organizations and low-resource political candidates or activists generate polished materials quickly, ranging from press releases to multilingual outreach. This shift could expand participation and coalition-building by lowering the costs of adaptation (Tomić et al. 2023; Kerley et al. 2024). Open-source intelligence (OSINT) research emphasizes that AI can help process large volumes of public information for extraction and knowledge discovery, which can support verification and accountability

PROOFS

work (Evangelista et al. 2021). Human rights practitioners argue for "authenticity infrastructure" and resilient witnessing practices to fortify the evidentiary value of digital media in an era of deepfakes (Gregory 2023). More broadly, scholarship on AI and content moderation stresses that algorithmic systems can be used either to suppress speech or to help create conditions for robust democratic exchange, depending on governance, accountability, and institutional incentives (Llansó 2020). Recent work on democracy movements similarly emphasizes tactical uses of AI for counter-messaging and real-time logistical support in contentious politics (Chenoweth 2025). In this way, while LLMs do appear to be facilitating greater digital repression and surveillance, they are also accelerating an adaptation race in which states, challengers, platforms, and civil society organizations continually update both offensive and defensive techniques to achieve their political goals (Feldstein 2019; Gregory 2023; Llansó 2020; Chenoweth 2025).

2.4 NEW APPROACHES TO RESEARCH

Generative AI may have not only important impacts at the micro and macro levels but also the power to change social science research on these topics. For instance, traditional experimental research designs that use static stimuli may be insufficient for testing theories about information exposure or understanding the policy implications of those findings. However, the relevant stimuli in the real world (e.g., political ads or news articles, exchanges with AI models) often vary in both systematic and idiosyncratic ways that are not captured by a single researcher-chosen stimulus, which often trades off ecological validity for analytical ease (Clifford, Leeper, and Rainey 2024; Clifford and Rainey 2024). This gap between research and practice may increase as politics, marketing, medicine, and education increasingly embrace personalization. Adaptive research designs are needed that reflect participants' information diets, backgrounds, and preferences. By adaptive, we mean research designs that tailor treatment or measurement to individual participants based on their characteristics, prior responses, or other contextual factors.

Current research on social media algorithms is already pushing in this direction. As public concern about the polarizing effects of algorithms has intensified, scholars have collaborated with social media companies to assess the political effects of recommendation algorithms, which vary at the individual level. As discussed earlier, Guess et al. (2023) randomly assign Facebook and Instagram users to either a default algorithm or a reverse-chronological feed. In contrast to traditional survey experiments employing fixed vignettes, the content of each feed varies at the individual level. This design recovers a policy effect rather than the effect of a single experimental stimulus.

Newer designs also leverage personalized stimuli to study attitudinal backfire (Velez and Liu 2025), the relationship between beliefs and attitudes (Velez, Liu, and Clifford 2026), and corrections of conspiracy beliefs

PROOFS

(Costello et al. 2024). For example, Velez and Liu (2025) use "tailored experiments" to assess whether motivated reasoning is activated when voters' deeply held issue stances are at stake. While backfire is rare in the literature (e.g., Coppock 2022), Velez and Liu (2025) use OpenAI's GPT-3 to generate personalized counterarguments targeting political convictions disclosed in an open-ended question. Across five studies, they find more evidence of moderation than polarization, with reliable evidence of backfire only emerging when counterarguments are presented using a vitriolic tone. This ability to align interventions with theoretically optimal conditions becomes possible with generative AI, which makes it possible for high-quality stimuli to be generated on the fly with little researcher supervision.

Tailored designs powered by LLMs can also enable adaptive measurement of rare or unexpected beliefs and attitudes. For example, Velez (2025) develops a crowdsourced adaptive survey method (CSAS) to convert open-ended responses from participants about misinformation and issue priorities into properly formatted survey questions using LLMs such as OpenAI's GPT-3 and GPT-4, leveraging multi-armed bandit algorithms to prioritize the most promising user-submitted questions. This method allows survey batteries to adapt to participants and surfaces new questions that might escape the notice of survey researchers.

Tailored approaches also offer a new way to test canonical theories. Many models in political science and psychology are built around inherently heterogeneous constructs. In theories of motivated reasoning, for example, strong issue attitudes are expected to activate defensiveness in response to counter-attitudinal information (Taber and Lodge 2006). Proximity voting models similarly emphasize "issue weights" as voters choose between candidates (e.g., Grynaviski and Corrigan 2006). Yet in both cases, researchers often operationalize these constructs using fixed issue batteries that feature a subset of policies. But if only a small share of the population is "activated" by the interventions (i.e., has a strongly held, stable attitude about the issue in question), such studies may be underpowered to detect theoretically meaningful relationships.

To provide an example, Velez, Liu, and Clifford (2026) use tailored experiments to resolve a long-standing puzzle about why belief change seldom produces attitude change. While numerous studies show positive effects of information on factual beliefs (Wood and Porter 2019; Hopkins, Sides, and Citrin 2019; Nyhan, Porter, Reifler, and Wood 2020), attitudes and behavior are remarkably resistant. Velez, Liu, and Clifford propose that the failure of counterarguments to shift attitudes hinges on "belief relevance" – the extent to which beliefs bear on attitudes. In a conversation with OpenAI's GPT-4o, participants are encouraged to provide reasons for their issue position ("focal beliefs"). The LLM then summarizes this justification, generates unmentioned justifications that support the participant's views ("distal beliefs"), and constructs counterarguments targeting both along with a placebo argument. The authors find that counterarguments targeting focal beliefs outperform those

PROOFS

targeting distal beliefs in producing attitude change, suggesting that belief relevance may condition the effects of counter-attitudinal information on attitudes.

Despite their promise, the move from fixed to tailored stimuli introduces a unique set of challenges. Stimuli such as text are already high-dimensional and personalization adds an additional layer of complexity. Scholars will need to carefully design their studies to recover the effects of bundled treatment consisting of model versions, parameters, participant inputs, and user interfaces. Such studies are not without precedent. Tailored experiments are not dissimilar from experiments that hinge on dynamic stimuli such as peer-to-peer conversations, where instantiations of the treatment vary across people. Still, it is worth recognizing that the mapping between this bundled treatment and the latent theoretical construct requires the (often strong) assumptions that the intended treatment is being manipulated and no stimulus-level confounding is present.

Careful prompting provides a path to manipulating latent constructs while minimizing known sources of confounding like length or readability. For example, researchers can guide an LLM to generate a counterargument of roughly the same length, valence, or tone across arms. Moreover, researchers can surgically vary certain instructions or remove them to assess potential mechanisms (e.g., Costello et al. 2025c; Argyle et al. 2025). However, while these approaches minimize known sources of confounding, LLMs might inherently bundle certain features in their internal representations of the text. For example, LLM-generated arguments about immigration may systematically co-vary with a "nationalism" feature, rendering it difficult to identify which text feature is producing the observed effect.

Imai and Nakamura (2026) propose a method of text deconfounding that leverages the internal representations of LLMs to minimize latent unobserved text confounding. This method exploits the fact that internal representations provide information about how LLMs encode semantic dimensions of generated text. Obtaining this internal representation requires accessing a model with open weights and extracting the hidden state from the final model layer before text generation. The method then estimates two models via double machine learning: a deconfounding model that predicts treatment status using latent text features and an outcome model that predicts outcomes using treatment status and these covariates. Conditioning on latent text features provides stronger evidence that estimated effects are not driven by unobserved features that covary with the treatment of interest (see also Roberts, Stewart, and Nielsen 2020).

Beyond text confounding, there is also the risk of "model drift." As Barrie, Palmer, and Spirling (2025) demonstrate, reproducibility challenges emerge when using proprietary LLMs that do not fix model weights. If papers use LLMs to classify text, and those LLMs are subsequently deprecated or model weights are modified, it becomes impossible to computationally reproduce classification procedures. Indeed, the authors discuss a number of models that entered a deprecated state after data collection. The authors caution against the use of proprietary models in social science research, advocating for "locally

PROOFS

versioned open models" that possess fixed weights and can be downloaded onto consumer devices.[1]

Despite these reproducibility challenges, use of proprietary models may be preferable if the interest is in studying how generative AI affects information seeking and persuasion in ecologically valid settings, such as elections, wars, or routine use by citizens. Given that political actors are likely to depend on high-performing models, assessing the persuasive capabilities and political biases of lesser-known, less effective open weights models risks addressing the wrong research question. If the goal is to inform public discourse about LLMs, prioritizing reproducibility over how political actors use these tools in real-world situations may limit the applicability of political science research. Proprietary models may also be preferred over open-source models if they are more effective in accomplishing research objectives including not just classification accuracy but, for instance, experimental designs where a strong "first-stage" relationship between the experimental manipulation and theoretical construct is critical for a fair test of the theory. If an open weights model is not capable of generating relevant stimuli, such a test will be less informative. Finally, it is worth recognizing that many real-world political constructs are stochastic and thus not completely reproducible. It is impossible to faithfully reproduce every aspect of a deliberation experiment, for instance, since the dynamics of conversations hinge on group composition and emergent interactions among participants. Large language model interventions may be similarly classified as "stochastic latent treatments" that cannot be held perfectly constant across replications.

We also anticipate generative AI models themselves becoming subjects of research. A long tradition in computer science has evaluated fairness, bias, and alignment to ensure that models are producing helpful and harmless outputs. Though several providers have undertaken audits of political bias, these audits depart from how political scientists tend to think about this construct. For example, OpenAI's audit of their models focused on dimensions of bias related to validation or invalidation of political beliefs, refusals, and whether political content was balanced (OpenAI 2025). Though these are sensible dimensions from a user experience perspective, they depart from the measurement models that have characterized the literature on ideology. Aldahoul et al. (2025) use item response models to measure political ideology across different LLMs, whereas Westwood, Grimmer, and Hall (2025) present participants with pairwise comparisons of model outputs and ask them to gauge bias. Such studies aim to measure political positions on a continuum, better approximating political bias than the "balanced responding" dimension that existing audits have captured.

[1] With few exceptions, such as Ai2's Olmo model, most "open source" LLMs, such as Meta's Llama, OpenAI's gpt-oss, and Google's Gemma, release model weights without training instructions or training data, departing from standard definitions of open source. The principal advantages of these models are that model weights are fixed and can be inspected.

PROOFS

Despite their advantages as research tools and subjects of research, generative AI models raise distinct ethical challenges. Chief among them is "hallucination," or the lack of factual grounding in model outputs. Though newer models have improved significantly by drawing on web searches, the quality of LLM outputs is uneven, with some referring to LLMs as possessing a "jagged intelligence," excelling at difficult tasks while inexplicably failing at simple ones (Karpathy 2024). While models may further improve, there are alternative architectures, such as RAG, that can improve factual responses. Retrieval-augmented generation supplements the responses of LLMs to user queries with information drawn from curated external databases. For example, Velez, Green, and Sevi (2025) use RAG to develop an AI-powered voting advice application that provides young independent voters with verified information about political parties. Though hallucinations may still persist (Magesh et al. 2025), such approaches – including newer agentic search techniques that retrieve and synthesize information from structured databases – can reduce the likelihood of unsupported or fabricated claims, while retaining the interactive qualities of generative AI.

Another ethical dimension lies in the research transparency and reproducibility challenges mentioned earlier. Demonstrating the capabilities of proprietary LLMs may be perceived as serving the commercial interests of their providers. However, avoiding the systematic study of proprietary models creates a vacuum that can be filled only by the providers themselves, who are trying their hand at estimating political bias in ways that deviate from how social scientists measure these concepts. Social scientists assessing the persuasive impact of proprietary models can help temper warnings of LLMs as hyper-persuasive agents. Such warnings likely do more to encourage malicious use than carefully conducted studies that provide precise effect sizes.

Though proprietary LLMs are often characterized as "black boxes" that lack transparency, it is worth comparing them against other information sources, such as social media platforms. In contrast to most social media algorithms, LLM application programming interfaces (or APIs) are publicly available, allowing researchers to probe how they respond to different queries and measure their behavior over time. For example, researchers can measure how models respond to contested political issues, job resumes of applicants from protected backgrounds, and other socially relevant queries. While the underlying weights are not available, public APIs allow for interventions and analyses that are not typically possible with social media algorithms absent bespoke tools developed by researchers (Piccardi et al. 2025). Compared to other algorithms shaping information environments, LLMs are unusually amenable to systematic study. That said, broader reproducibility and transparency concerns would be significantly addressed if LLM providers were to open-source more models and "freeze" specific versions for academic research, committing to maintain them over time.

PROOFS

2.5 CONCLUSION

This review suggests that the effects of AI on persuasion and the online information ecosystem will be more subtle than many observers expect. Artificial intelligence will change what news and information people consume substantially, in part by displacing search engines from their key role in online information acquisition. But differences in information exposure are unlikely to transform mass opinion, which is constrained by public inattention and preexisting attitudes. The most substantial effects of AI could be the ways in which it transforms the capabilities of platforms, state actors, and nonstate actors to engage in content production, monitoring, and surveillance at unprecedented scale, especially in authoritarian regimes. But large-scale targeted harassment and "flooding" attacks on information spaces are now cheap and easy to conduct even for individuals. Importantly, these tactics can all be effective without changing minds.

We caution that these conclusions apply unevenly throughout the world and may change over time. In particular, second-order effects of AI ubiquity on trust in online information and the effectiveness of "liar's dividend"-style denials of unwelcome fact could dominate AI's first-order effects on what people see and believe. For this reason, it is essential to leverage all the research tools available to us, including AI, to understand the effects it is having on the world.

REFERENCES

Aldahoul, N., Ibrahim, H., Varvello, M., Kaufman, A., Rahwan, T., & Zaki, Y. (2025). *Large language models are often politically extreme, usually ideologically inconsistent, and persuasive even in informational contexts*. arXiv [preprint] arXiv:2505.04171. https://doi.org/10.48550/arXiv.2505.04171

Altay, S., & Gilardi, F. (2024). People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation. *PNAS Nexus*, *3*(10), pgae403. https://doi.org/10.1093/pnasnexus/pgae403

Altman, S. (2023, May 16). Testimony before the U.S. Senate Committee on the Judiciary, Subcommittee on Privacy, Technology, and the Law. Hearing: "Oversight of A.I.: Rules for Artificial Intelligence." 118th Congress. www.congress.gov/event/118th-congress/senate-event/LC71543/text

Alyukov, M., Makhortykh, M., Voronovici, A., & Sydorova, M. (2025). LLMs grooming or data voids? LLM-powered chatbot references to Kremlin disinformation reflect information gaps, not manipulation. *Harvard Kennedy School Misinformation Review*, *6*(5). https://doi.org/10.37016/mr-2020-187

Amodei, D. (2026, January 26). The adolescence of technology: Confronting and overcoming the risks of powerful AI. www.darioamodei.com/essay/the-adolescence-of-technology

Aral, S., Li, H., & Zuo, R. (2026). *The rise of AI search: Implications for information markets and human judgement at scale*. arXiv [preprint] arXiv:2602.13415.

PROOFS

Arguedas, A. R., Robertson, C. T., Fletcher, R., & Nielsen, R. K. (2022). Echo chambers, filter bubbles, and polarisation: A literature review. Reuters Institute for the Study of Journalism, University of Oxford. https://doi.org/10.60625/risj-etxj-7k60

Argyle, L. P., Busby, E. C., Gubler, J. R., Lyman, A., Olcott, J., Pond, J., & Wingate, D. (2025). Testing theories of political persuasion using AI. *Proceedings of the National Academy of Sciences*, *122*(18), e2412815122. https://doi.org/10.1073/pnas.2412815122

Bai, H., Voelkel, J. G., Muldowney, S., Eichstaedt, J. C., & Willer, R. (2025). LLM-generated messages can persuade humans on policy issues. *Nature Communications*, *16*, 6037. https://doi.org/10.1038/s41467-025-61345-5

Barrie, C., Palmer, A., & Spirling, A. (2025). Replication for language models: Problems, principles, and best practice for political science. https://arthurspirling.org/documents/BarriePalmerSpirling_TrustMeBro.pdf

Blair, R. A., Gottlieb, J., Nyhan, B., Paler, L., Argote, P., & Stainfield, C. J. (2024). Interventions to counter misinformation: Lessons from the Global North and applications to the Global South. *Current Opinion in Psychology*, *55*, 101732. https://doi.org/10.1016/j.copsyc.2023.101732

Blumenthal, R. (2024, April 17). ICYMI video: Blumenthal highlights dangers of artificial intelligence in spreading disinformation with election deepfakes [Press release]. U.S. Senate, Subcommittee on Privacy, Technology, and the Law. https://tinyurl.com/46xe2xcm

Boissin, E., Costello, T. H., Spinoza-Martín, D., Rand, D. G., & Pennycook, G. (2025). Dialogues with large language models reduce conspiracy beliefs even when the AI is perceived as human. *PNAS Nexus*, *4*(11), pgaf325. https://doi.org/10.1093/pnasnexus/pgaf325

Bond, S., & Brumfiel, G. (2025, December 24). 2025 has seen an explosion of AI-generated slop. *NPR*. www.npr.org/2025/12/24/nx-s1-5629169/2025-has-seen-an-explosion-of-ai-generated-slop

Bradshaw, S., & Howard, P. N. (2017). Troops, trolls and troublemakers: A global inventory of organized social media manipulation (Working Paper 2017.12). Computational Propaganda Research Project, Oxford Internet Institute, University of Oxford. https://ora.ox.ac.uk/objects/uuid:cef7e8d9-27bf-4ea5-9fd6-855209b3e1f6

Budak, C., Nyhan, B., Rothschild, D. M., Thorson, E., & Watts, D. J. (2024). Misunderstanding the harms of online misinformation. *Nature*, *630*(8015), 45–53. https://doi.org/10.1038/s41586-024-07417-w

Campante, F. R., Durante, R., Hagemeister, F., & Sen, A. (2025). GenAI misinformation, trust, and news consumption: Evidence from a field experiment (NBER Working Paper No. 34100). National Bureau of Economic Research. www.nber.org/papers/w34100

Carnegie Endowment for International Peace [CEIP]. (2019, September). The global expansion of AI surveillance (S. Feldstein, Author). https://tinyurl.com/4tczy67z

Châtelet, V. (2025, April 18). Exposing Pravda: How pro-Kremlin forces are poisoning AI models and rewriting Wikipedia. Atlantic Council, Digital Forensic Research Lab. https://tinyurl.com/mre6zb3a

PROOFS

Chatterji, A., Cunningham, T., Deming, D., Hitzig, Z., Ong, C., Shan, C. Y., & Wadman, K. (2025). How people use ChatGPT (NBER Working Paper No. 34255). National Bureau of Economic Research. https://doi.org/10.3386/w34255

Chenoweth, E. (2025, February). How AI can support democracy movements: Summary report of a research and practice workshop. Ash Center for Democratic Governance and Innovation, Harvard Kennedy School. https://tinyurl.com/mat5tff

Clifford, S., Leeper, T. J., & Rainey, C. (2024). Generalizing survey experiments using topic sampling: An application to party cues. *Political Behavior*, *46*, 1233–1256. https://doi.org/10.1007/s11109-023-09870-1

Clifford, S., & Rainey, C. (2024). Estimators for topic-sampling designs. *Political Analysis*, *32*(4), 431–444. https://doi.org/10.1017/pan.2024.1

Coppock, A. (2022). *Persuasion in parallel: How information changes minds about politics*. University of Chicago Press.

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

Costello, T. H., Rabb, N., Stagnaro, M. N., Pennycook, G., & Rand, D. G. (2025a). *Reducing belief in conspiracy theories as they unfold using large language models*. PsyArXiv [preprint]. https://doi.org/10.31234/osf.io/jnm9r_v1

Costello, T. H., Pennycook, G., Willer, R., & Rand, D. G. (2025b). *Deep canvassing using AI*. OSF [preprint]. https://doi.org/10.31219/osf.io/q7e6u_v1

Costello, T. H., Pennycook, G., & Rand, D. G. (2025c). *Just the facts: How dialogues with AI reduce conspiracy beliefs*. PsyArXiv [preprint]. https://doi.org/10.31234/osf.io/h7n8u_v1

Costello, T. H., Pelrine, K., Kowal, M., Arechar, A. A., Godbout, J.-F., Gleave, A., Rand, D., & Pennycook, G. (2026). *Large language models can effectively convince people to believe conspiracies*. arXiv [preprint] arXiv:2601.05050. https://doi.org/10.48550/arXiv.2601.05050

Cox, J. (2026, February 12). Cops are buying GeoSpy, an AI that geolocates photos in seconds. *404 Media*. www.404media.co/cops-are-buying-geospy-ai-that-geolocates-photos-in-seconds/

Czarnek, G., Orchinik, R., Lin, H., Xu, H. G., Costello, T., Pennycook, G., & Rand, D. G. (2025). *Addressing climate change skepticism and inaction using human-AI dialogues*. PsyArXiv [preprint]. https://doi.org/10.31234/osf.io/mqcwj_v1

Danet, D. (2025). LLM grooming: A new cognitive threat to generative AI. Working paper, Centre Géopolitique de la Datasphère. HAL: hal-05241525. https://hal.science/hal-05241525/

del Rio-Chanona, R. M., Laurentsyeva, N., & Wachs, J. (2024). Large language models reduce public knowledge sharing on online Q&A platforms. *PNAS Nexus*, *3*(9), pgae400. https://doi.org/10.1093/pnasnexus/pgae400

Deng, Y., Zhang, W., Pan, S. J., & Bing, L. (2024). *Multilingual jailbreak challenges in large language models*. arXiv [preprint] arXiv:2310.06474.

DeVerna, M. R., Yan, H. Y., Yang, K.-C., & Menczer, F. (2024). Fact-checking information from large language models can decrease headline discernment. *Proceedings of the National Academy of Sciences*, *121*(50), e2322823121. https://doi.org/10.1073/pnas.2322823121

PROOFS

Drotner, K. (1999). Dangerous media? Panic discourses and dilemmas of modernity. *Paedagogica Historica*, *35*(3), 593–619. https://doi.org/10.1080/0030923990350303

Espiritusanto, O., Nachawati-Rego, L., & Magallón-Rosa, R. (2024). The role of AI in citizen journalism, human rights activism, and monitoring: Limits and possibilities. In J. Sixto-García, A. Quian, A.-I. Rodríguez-Vázquez, A. Silva-Rodríguez, & X. Soengas-Pérez (Eds.), *Journalism, digital media and the fourth industrial revolution* (pp. 211–226). Springer. https://doi.org/10.1007/978-3-031-63153-5_16

Evangelista, J. R. G., Sassi, R. J., Romero, M., & Napolitano, D. (2021). Systematic literature review to investigate the application of open source intelligence (OSINT) with artificial intelligence. *Journal of Applied Security Research*, *16*(3), 345–369. https://doi.org/10.1080/19361610.2020.1761737

Feldstein, S. (2019). The road to digital unfreedom: How artificial intelligence is reshaping repression. *Journal of Democracy*, *30*(1), 40–52.

Fisher, J., Appel, R. E., Park, C. Y., Potter, Y., Jiang, L., Sorensen, T., Feng, S., Tsvetkov, Y., Roberts, M. E., Pan, J., Song, D., & Choi, Y. (2025a). *Political neutrality in AI is impossible – but here is how to approximate it*. arXiv [preprint] arXiv:2503.05728.

Fisher, J., Feng, S., Aron, R., Richardson, T., Choi, Y., Fisher, D. W., Pan, J., Tsvetkov, Y., & Reinecke, K. (2025b). Biased LLMs can influence political decision-making. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 6559–6607). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.328

Freedom House. (2023). *Freedom on the Net 2023: The repressive power of artificial intelligence*. https://tinyurl.com/3k8eh23k

Gallegos, I. O., Shani, C., Shi, W., Bianchi, F., Gainsburg, I., Jurafsky, D., & Willer, R. (2025). *Labeling messages as AI-generated does not reduce their persuasive effects*. arXiv [preprint] arXiv:2504.09865.

Goldstein, J. A., Chao, J., Grossman, S., Stamos, A., & Tomz, M. (2024). How persuasive is AI-generated propaganda? *PNAS Nexus*, *3*(2), pgae034. https://doi.org/10.1093/pnasnexus/pgae034

Golebiewski, M., & Boyd, D. (2018). Data voids: Where missing data can easily be exploited. Data & Society Research Institute. https://tinyurl.com/52pdrhhb

Gregory, S. (2023). Fortify the truth: How to defend human rights in an age of deepfakes and generative AI. *Journal of Human Rights Practice*, *15*(3), 702–714. https://doi.org/10.1093/jhuman/huad035

Grynaviski, J. D., & Corrigan, B. E. (2006). Specification issues in proximity models of candidate evaluation (with issue importance). *Political Analysis*, *14*(4), 393–420. https://doi.org/10.1093/pan/mpl003

Guess, A. M. (2021). (Almost) everything in moderation: New evidence on Americans' online media diets. *American Journal of Political Science*, *65*(4), 1007–1022. https://doi.org/10.1111/ajps.12589

Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., … & Tucker, J. A. (2023). How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, *381*(6656), 398–404. https://doi.org/10.1126/science.abp9364

Guterres, A. (2023, July 18). Secretary-General urges Security Council to ensure transparency, accountability, oversight, in first debate on artificial intelligence [Remarks to the UN Security Council]. United Nations. https://tinyurl.com/2nhhfjh7

PROOFS

Hackenburg, K., & Margetts, H. (2024). Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences*, *121*(24), e2403116121. https://doi.org/10.1073/pnas.2403116121

Hackenburg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D. G., & Summerfield, C. (2025). The levers of political persuasion with conversational artificial intelligence. *Science*, *390*(6777), eaea3884. https://doi.org/10.1126/science.aea3884

Hopkins, D. J., Sides, J., & Citrin, J. (2019). The muted consequences of correct information about immigration. *Journal of Politics*, *81*(1), 315–320. https://doi.org/10.1086/699914

Hubbard, S. (2025, December 18). ICE uses a growing web of AI services to power its immigration enforcement and surveillance. American Immigration Council. https://tinyurl.com/4w5z4xf9

Imai, K., & Nakamura, K. (2026). Causal inference with generative artificial intelligence: Application to texts as treatments. *Journal of the American Statistical Association*. Advance online publication. https://doi.org/10.1080/01621459.2026.2689629.

Jingnan, H. (2024, October 18). AI-generated images have become a new form of propaganda this election season. *NPR*. www.npr.org/2024/10/18/nx-s1-5153741/ai-images-hurricanes-disasters-propaganda

Karpathy, A. [@karpathy]. (2024, July 25). Jagged intelligence [Post]. X (formerly Twitter). https://x.com/karpathy/status/1816531576228053133

Kerley, B., Miller, C., & Campagnucci, F. (2024). Leveraging AI for democracy. In National Endowment for Democracy's International Forum for Democratic Studies. www.ned.org/wp-content/uploads/2024/10/NED_Leveraging-AI-for-Democracy-Report.pdf

Kowal, M., Timm, J., Godbout, J.-F., Costello, T., Arechar, A. A., Pennycook, G., Rand, D., Gleave, A., & Pelrine, K. (2025). *It's the thought that counts: Evaluating the attempts of frontier LLMs to persuade on harmful topics*. arXiv [preprint] arXiv:2506.02873.

Lawler, D., & Curi, M. (2026, February 15). Exclusive: Pentagon threatens to cut off Anthropic in AI safeguards dispute. *Axios*. www.axios.com/2026/02/15/claude-pentagon-anthropic-contract-maduro

Li, Z., Shi, Y., Liu, Z., Yang, F., Payani, A., Liu, N., & Du, M. (2025). Language ranker: A metric for quantifying LLM performance across high and low-resource languages. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(27), 28186–28194.

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

Llansó, E. J. (2020). No amount of "AI" in content moderation will solve filtering's prior-restraint problem. *Big Data & Society*, *7*(1). https://doi.org/10.1177/2053951720920686

Lu, L., Tormala, Z. L., & Duhachek, A. (2025). How AI sources can increase openness to opposing views. *Scientific Reports*, *15*, 17170. https://doi.org/10.1038/s41598-025-00791-z

Lum, D., & Willis, H. (2026). Videos show moments in which agents killed a man in Minneapolis. *New York Times*, January 24, 2026. www.nytimes.com/2026/01/24/us/minneapolis-shooting-federal-agents-video.html

PROOFS

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). Hallucination-free? Assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies*, *22*(2), 216–242. https://doi.org/10.1111/jels.12413

Marcellino, W., Beauchamp-Mustafaga, N., Kerrigan, A., Chao, L. N., & Smith, J. (2023). *The rise of generative AI and the coming era of social media manipulation 3.0*. RAND Corporation, PE-A2679-1. www.rand.org/pubs/perspectives/PEA2679-1.html

Matz, S. C., Teeny, J. D., Vaid, S. S., Peters, H., Harari, G. M., & Cerf, M. (2024). The potential of generative AI for personalized persuasion at scale. *Scientific Reports*, *14*, 4692. https://doi.org/10.1038/s41598-024-53755-0

McGuffie, K., & Newhouse, A. (2020). *The radicalization risks of GPT-3 and advanced neural language models*. arXiv [preprint] arXiv:2009.06807.

Meta. (2025, October 7). Improving your recommendations on our apps with AI at Meta. *Meta Newsroom*. https://about.fb.com/news/2025/10/improving-your-recommendations-apps-ai-meta/

Motoki, F., Pinho Neto, V., & Rangel, V. (2025). Assessing political bias and value misalignment in generative artificial intelligence. *Journal of Economic Behavior & Organization*, *234*, 106904. https://doi.org/10.1016/j.jebo.2025.106904

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

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

Nanz, A., Binder, A., & Matthes, J. (2025). AI in the newsroom: Does the public trust automated journalism and will they pay for it? *Journalism Studies*, *26*(14), 1745–1764. https://doi.org/10.1080/1461670X.2025.2547301

Noels, S., Bied, G., Buyl, M., Rogiers, A., Fettach, Y., Lijffijt, J., & De Bie, T. (2025). What large language models do not talk about: An empirical study of moderation and censorship practices. In Machine Learning and Knowledge Discovery in Databases: Research Track, ECML PKDD 2025, Lecture Notes in Computer Science (vol. 16013). Springer. https://doi.org/10.1007/978-3-032-05962-8_16

Nyhan, B., Porter, E., Reifler, J., & Wood, T. J. (2020). Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political Behavior*, *42*(3), 939–960. https://doi.org/10.1007/s11109-019-09528-x

OpenAI. (2025, October 9). Defining and evaluating political bias in LLMs. https://openai.com/index/defining-and-evaluating-political-bias-in-llms/

Orben, A. (2020). The Sisyphean cycle of technology panics. *Perspectives on Psychological Science*, *15*(5), 1143–1157. https://doi.org/10.1177/1745691620919372

Pan, J., & Xu, X. (2026). Political censorship in large language models originating from China. *PNAS Nexus*, *5*(2), pgag013. https://doi.org/10.1093/pnasnexus/pgag013

Papathanassopoulos, S., & Giannouli, I. (2025). Political communication in the age of platforms. *Encyclopedia*, *5*(2), 77. https://doi.org/10.3390/encyclopedia5020077

Park, J., & Kim, J. (2024, March 18). South Korea's Yoon warns of tech threat to democracy at summit. *Reuters*. https://tinyurl.com/2tp6shfr

Pauwels, E. (2024). *Preparing for next-generation information warfare with generative AI* (CIGI Paper No. 310). Centre for International Governance Innovation. www.econstor.eu/handle/10419/311791

PROOFS

Pava, J. N., Meinhardt, C., Zaman, H. B. U., Friedman, T., Truong, S. T., Zhang, D., Marivate, V., & Koyejo, S. (2025). "Mind the (language) gap: Mapping the challenges of LLM development in low-resource language contexts." In *Stanford Institute for Human-Centered Artificial Intelligence (HAI): Stanford, CA, USA*.

Piccardi, T., Saveski, M., Jia, C., Hancock, J., Tsai, J. L., & Bernstein, M. S. (2025). Reranking partisan animosity in algorithmic social media feeds alters affective polarization. *Science*, *390*(6776), eadu5584. https://doi.org/10.1126/science.adu5584

Potter, Y., Lai, S., Kim, J., Evans, J., & Song, D. (2024). Hidden persuaders: LLMs' political leaning and their influence on voters. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*. arXiv:2410.24190.

Puczyńska, J., Podhajski, M., Wojtasik, K., & Michalak, T. P. (2024). Large language models in jihadist terrorism and crimes. *Terrorism – Studies, Analyses, Prevention*, (5), 351–379. https://doi.org/10.4467/27204383TER.24.012.19400

Qi, T., Liu, H., & Huang, Z. (2025). An assistant or a friend? The role of parasocial relationship of human-computer interaction. *Computers in Human Behavior*, *167*, 108625. https://doi.org/10.1016/j.chb.2025.108625

Qiu, P., Zhou, S., & Ferrara, E. (2025). *Information suppression in large language models: Auditing, quantifying, and characterizing censorship in DeepSeek*. arXiv [preprint] arXiv:2506.12349.

Quinn, M., & Gutt, D. (2025). Heterogeneous effects of generative artificial intelligence (GenAI) on knowledge seeking in online communities. *Journal of Management Information Systems*, *42*(2), 370–399. https://doi.org/10.1080/07421222.2025.2487313

Roberts, M. E., Stewart, B. M., & Nielsen, R. A. (2020). Adjusting for confounding with text matching. *American Journal of Political Science*, *64*(4), 887–903.

Robinson, N., Ogayo, P., Mortensen, D. R., & Neubig, G. (2023). ChatGPT MT: Competitive for high- (but not low-) resource languages. In Proceedings of the Eighth Conference on Machine Translation (pp. 392–418). Association for Computational Linguistics. https://aclanthology.org/2023.wmt-1.40.pdf

Rozado, D. (2024). The political preferences of LLMs. *PLOS ONE*, *19*(7), e0306621. https://doi.org/10.1371/journal.pone.0306621

Rozado, D. (2025). *Measuring political preferences in AI systems: An integrative approach*. arXiv [preprint] arXiv:2503.10649.

Salvi, F., Horta Ribeiro, M., Gallotti, R., & West, R. (2025). On the conversational persuasiveness of GPT-4. *Nature Human Behaviour*, *9*(8), 1645–1653. https://doi.org/10.1038/s41562-025-02194-6

Sanderson, Z., Messing, S., & Tucker, J.A. (2024). Misunderstood mechanics: How AI, TikTok, and the liar's dividend might affect the 2024 elections. Brookings Institution. https://tinyurl.com/ycyf58wf

Sanderson, Z., Zhong, W., & Tucker, J.A. (2025). It works when it works: Measuring the direct and indirect effects of AI labels on political images. https://osf.io/preprints/socarxiv/nf785_v1

Schiff, K. J., Schiff, D. S., & Bueno, N. S. (2025). The liar's dividend: Can politicians claim misinformation to evade accountability? *American Political Science Review*, *119*(1), 71–90. https://doi.org/10.1017/S0003055423001454

Schneier, B., & Sanders, N. (2024, December 2). The apocalypse that wasn't: AI was everywhere in 2024's elections, but deepfakes and misinformation were only part of

PROOFS

the picture. *The Conversation*. Republished by Harvard Ash Center. https://tinyurl.com/5xt8942b

Sen, R., & Farooq, N. (2025). AI-driven digital transnational repression: Past lessons, present challenges, and future directions. In M. Hasan & A. E. Ruud (Eds.), *The long reach of the strong arm: Evolving forms of transnational authoritarianism* (pp. 31–59). Palgrave Macmillan. https://doi.org/10.1007/978-3-032-04940-7_3

Setty, R. (2026). In this Cleveland Newsroom, AI is writing (but not reporting) the news. *Columbia Journalism Review*, February 25, 2026. www.cjr.org/news/cleveland-newsroom-ai-rewrite-desk-chris-quinn-plain-dealer.php

Shambaugh, S. (2026). An AI agent published a hit piece on me, February 12, 2026. https://theshamblog.com/an-ai-agent-published-a-hit-piece-on-me/

Sharma, N., Murray, K., & Xiao, Z. (2025). *Faux polyglot: A study on information disparity in multilingual large language models*. arXiv [preprint] arXiv:2407.05502.

Similarweb. (2025). *GenAI and how it's impacting US publishers*. Similarweb. www.similarweb.com/corp/reports/generative-ai-publishers/

Smith, M. G., Bradbury, T. N., & Karney, B. R. (2025). Can generative AI chatbots emulate human connection? A relationship science perspective. *Perspectives on Psychological Science*, *20*(6), 1081–1099. https://doi.org/10.1177/17456916251351306

Spearing, E. R., Gile, C. I., Fogwill, A. L., Prike, T., Swire-Thompson, B., Lewandowsky, S., & Ecker, U. K. H. (2025). Countering AI-generated misinformation with pre-emptive source discreditation and debunking. *Royal Society Open Science*, *12*(6), 242148. https://doi.org/10.1098/rsos.242148

Starker, S. (1989). *Evil influences: Crusades against the mass media*. Transaction Publishers.

Summerfield, C., Argyle, L. P., Bakker, M., et al. (2025). The impact of advanced AI systems on democracy. *Nature Human Behaviour*, *9*(12), 2420–2430. https://doi.org/10.1038/s41562-025-02309-z

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, *50*(3), 755–769. https://doi.org/10.1111/j.1540-5907.2006.00214.x

Thekkethil, D. (2025, July 7). Similarweb: Zero-click searches surge to 69% since Google AI Overviews launched. *Stan Ventures*. www.stanventures.com/news/similarweb-zero-click-search-surge-google-ai-overviews-3562/

TikTok. (2024). Our approach to content moderation. TikTok Transparency Center. www.tiktok.com/transparency/en-us/content-moderation

Tomić, Z., Damnjanović, T., & Tomić, I. (2023). Artificial intelligence in political campaigns. *South Eastern European Journal of Communication*, *5*(2), 17–28. https://doi.org/10.47960/2712-0457.2.5.17

Trump, D. J. (2025). Preventing woke AI in the federal government. White House executive order, July 23, 2025. https://tinyurl.com/48rj2p2j

Velez, Y. R. (2025). Crowdsourced adaptive surveys. *Political Analysis*, *33*(4), 284–297. https://doi.org/10.1017/pan.2024.34

Velez, Y. R., & Liu, P. (2025). Confronting core issues: A critical assessment of attitude polarization using tailored experiments. *American Political Science Review*, *119*(2), 1036–1053. https://doi.org/10.1017/S0003055424000819

PROOFS

Velez, Y. R., Green, D. P., & Sevi, S. (2025). Chatbot voting advice applications inform but seldom sway young unaligned voters. *Proceedings of the National Academy of Sciences*, *122*(50), e2515516122. https://doi.org/10.1073/pnas.2515516122

Velez, Y. R., Liu, P., & Clifford, S. (2026). When information affects attitudes: The effectiveness of targeting attitude-relevant beliefs. APSA Preprints. https://preprints.apsanet.org/engage/apsa/article-details/67e7f3f981d2151a02322745

Wack, M., Ehrett, C., Linvill, D., & Warren, P. (2025). Generative propaganda: Evidence of AI's impact from a state-backed disinformation campaign. *PNAS Nexus*, *4*(4), pgaf083. https://doi.org/10.1093/pnasnexus/pgaf083

Waight, H., Yang, E., Yuan, Y., Messing, S., Roberts, M., Stewart, B., & Tucker, J. (2026). State media control influences large language models. *Nature*. Advance online publication. https://doi.org/10.1038/s41586-026-10506-7.

Westwood, S. J., Grimmer, J., & Hall, A. B. (2025). Measuring perceived slant in large language models through user evaluations (Working Paper No. 4262). Stanford Graduate School of Business. https://tinyurl.com/yc394dhn

Williams, A. R., Burke-Moore, L., Chan, R. S.-Y., et al. (2025). Large language models can consistently generate high-quality content for election disinformation operations. *PLOS ONE*, *20*(3), e0317421. https://doi.org/10.1371/journal.pone.0317421

Wilson, A. 2026. Instagram algorithm 2026 – What changed, what works, what to stop doing. *OrangeMonke*. https://orangemonke.com/blogs/instagram-algorithm/

Wirtschafter, V., & Nadgir, N. (2025, October 16). Is the politicization of generative AI inevitable? Brookings Institution. www.brookings.edu/articles/is-the-politicization-of-generative-ai-inevitable/

Wittenberg, C., Epstein, Z., Péloquin-Skulski, G., Berinsky, A. J., & Rand, D. G. (2025). Labeling AI-generated media online. *PNAS Nexus*, *4*(6), pgaf170. https://doi.org/10.1093/pnasnexus/pgaf170

Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, *41*(1), 135–163. https://doi.org/10.1007/s11109-018-9443-y

Yang, E., & Roberts, M. E. (2023). The authoritarian data problem. *Journal of Democracy*, *34*(4), 141–150.

Yildiz, G. (2026). Inside Moltbook: The social network where AI agents talk and humans just watch. *Forbes*. https://tinyurl.com/yjkxsh96

Yong, Zheng-Xin, Cristina Menghini, & Stephen H. Bach. (2023). *Low-resource languages jailbreak GPT-4*. arXiv [preprint] arXiv:2310.02446.

Zhuravskaya, E., Petrova, M., & Enikolopov, R. (2020). Political effects of the internet and social media. *Annual Review of Economics*, *12*(1), 415–438. https://doi.org/10.1146/annurev-economics-081919-050239

PROOFS

3

Public Opinion in the Age of AI

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

PROOFS

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

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

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

Tucker_9781047774734_C003.indd 78 07-09-2026 12:37:28

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

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

3.1 THE CONCEPT OF "PUBLIC OPINION"

It is difficult to overstate the importance of public opinion to governing and governments. Even authoritarian regimes must attend to public sentiment, if only to avoid provoking instability or revolt. As V. O. Key famously observed, "Governments must concern themselves with the opinions of their citizens, if only to provide a basis for repression of disaffection" (Key, 1961). In representative democracies, public opinion occupies an even more central role, shaping debates over political responsiveness, accountability, and legitimacy. Longstanding normative disagreements persist over whether elected officials should exercise independent judgment or defer to constituent preferences (Burke, 1774), and some have argued that responsiveness to public opinion represents the highest form of democratic government (Bryce, 1888).

Despite its centrality, "public opinion" is a remarkably elusive concept. Scholars have defined it in multiple, often incompatible ways: as the aggregation of individual preferences (Key, 1961), the considered judgments of an informed and deliberative public (Fishkin, 1991), the expression of organized social groups (Blumer, 1948), or a mediated construction shaped by elites, institutions, and information environments (Lippmann, 1922). Even

PROOFS

abstracting from problems of measurement, classical results in social choice theory demonstrate that aggregating individual preferences may yield indeterminate or unstable collective outcomes, even when individual preferences are internally consistent (Arrow, 1951). At the individual level, many citizens lack well-formed or stable political attitudes, instead offering responses that reflect the wording of a question, contextual cues, or sampled considerations, rather than durable beliefs (Converse, 1964; Sniderman et al., 1991; Zaller, 1992). The conceptual foundations of public opinion are therefore unsettled even before empirical measurement enters the picture.

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

PROOFS

3.1.1 Technological Change and Public Opinion

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

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

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

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

PROOFS

If social media reshaped public opinion by altering the flow of social information, AI represents a potentially more profound shift in how individuals encounter, interpret, and generate politically relevant content. Social media primarily affects what people believe *others* think, but AI systems intervene more directly in processes of information seeking, judgment formation, and expression. Rather than merely mediating or reporting information, AI tools can generate content, personalize it at scale, and influence how information is weighted and surfaced by downstream systems. Artificial intelligence–generated bots may shape public opinion through targeted content or by affecting the signals that guide platform algorithms (Schroeder et al., 2026), and systems optimized for affirmation may reinforce inferred beliefs rather than promote updating or exposure to competing perspectives (Bisbee et al., 2026). Recent experiments confirm that AI chatbots can durably reduce conspiracy beliefs and shift voter preferences by substantial margins across multiple countries (Costello et al., 2024; Hackenburg and Margetts, 2024; Lin et al., 2025), and audits suggest that model outputs tend to lean more liberal and cosmopolitan than the average voter (Santurkar et al., 2023; Westwood et al., 2025; Lyman et al. 2025). While not the focus of our chapter, understanding how public opinion itself is shaped by people's encounters and usage of AI in ordinary contexts is a topic of increasing importance.

3.2 THE EFFECT(S) OF AI ON SURVEY-BASED MEASURES OF PUBLIC OPINION

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

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

Whereas the first use largely avoids conceptual claims about the epistemological status of AI outputs, imputation-based uses require treating AI-generated

PROOFS

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

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

Turning to the ways in which AI is used in survey methodology, Buskirk and colleagues (2025a) explore the uses of LLMs within three broad phases of the survey research process: tasks prior to data collection, data collection itself, and post -data-collection activities (see also Rothschild et al., 2026). Most uses currently involve data- collection and post -data-collection phases. Figure 3.1 reports examples that include *coding open-ended responses* (Mellon et al., 2024; Rytting et al., 2023; Singh and Kumar, 2025; von der Heyde et al., 2025), *interviewing* (Barari et al., 2025; Lang and Eskenazi, 2025; Wuttke et al., 2025; Xiao et al., 2020), *pretesting and survey development* (Buskirk et al., 2025b Adhikari et al., 2025; Tao et al., 2024; Yun et al., 2024), and *analyzing responses* (Ho et al., 2025 and Huang et al., 2024).

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

PROOFS

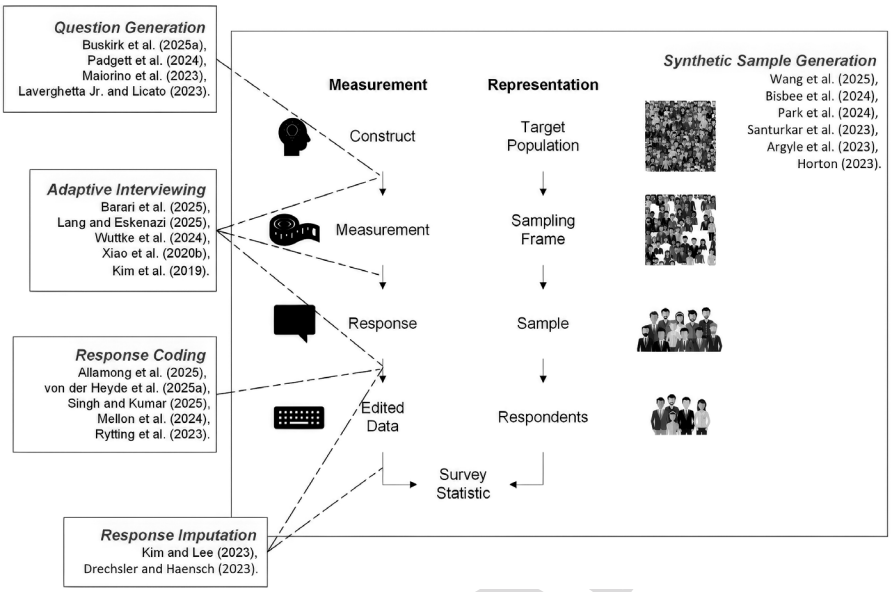


FIGURE 3.1 **Selected uses of AI in survey research.** Illustration shows some ways AI is currently used in survey research: Generating survey items, pretesting, AI-assisted interviews, generating synthetic survey responses and imputation, and coding open-ended responses.

AI into the Total Survey Error framework, but applying that framework going forward offers a structured way to evaluate how AI may both reduce and generate distinct forms of survey error.

3.2.1 Pre-data Collection

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

A growing line of research examines the use of LLMs for generating surveys at both the item and questionnaire level (e.g., Behrend and Landers, 2025). Empirical findings suggest that LLMs can produce relevant and comprehensible questions with response options that align closely with the intended topic (Maiorino et al., 2023; Padgett et al., 2024). However, question quality depends

PROOF

heavily on prompt design and input structure. Models tend to exhibit consistent stylistic regularities, such as preferences for binary or five-point Likert response formats, difficulties with causal precision, and challenges maintaining coherent ordering across longer instruments (Padgett et al., 2024; Buskirk et al., 2025b; Fuchs et al., 2026). Absent additional constraints, generated items are often written at relatively high reading levels, underscoring the need for independent evaluation using established survey quality metrics (Buskirk et al., 2025c; Olson and Buskirk, 2025). Overall, systematic evidence remains limited as to the measurement quality and construct validity of AI-generated survey items.

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

More expansive and interpretive uses of AI in pretesting that involve the simulation of respondent understanding or generating responses from hypothetical personas introduce greater inferential risks. Studies show that LLMs can flag potential design flaws or provide preliminary feedback prior to human pretesting, but performance depends on prompt design and the availability of best-practice examples (Tao et al., 2024; Olivos and Liu, 2024). When used to generate responses for exploratory analysis or theory development, such approaches raise concerns about representativeness, bias, and model-induced agreement, particularly when outputs are treated as informative about real populations rather than as heuristic inputs (Anthis et al., 2025; Park et al., 2026).

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

3.2.2 Data Collection

Artificial intelligence can also be incorporated in the data- collection phase of survey research, most prominently through AI-assisted interviewing. These approaches use AI models to conduct interactive, conversation-based interviews that tailor question wording, sequencing, and probing to respondents.

PROOFS

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

Artificial intelligence–assisted interviewing offers several potential methodological benefits, including the ability to clarify concepts to the respondent or interviewer while interviewing, real-time analysis of open-ended responses, and adaptive branching (Barari et al., 2025; Lang and Eskenazi, 2025). Personalization may improve probing of core issue positions or time-sensitive, high-salience topics (Velez and Liu, 2025; Velez, 2025), and some studies report that AI-assisted interviews elicit more informative and contextually relevant responses than traditional web surveys (Barari et al., 2025; Xiao et al., 2020). The use of AI may also reduce interviewer effects (West and Blom, 2017) and social desirability bias (Krumpal, 2013), increase engagement (Yun et al., 2024; Lang and Eskenazi, 2025), and offer advantages in scalability, cost, and operational efficiency (Wuttke et al., 2025). However, these benefits are design- and mode-dependent: Increased interactivity can also raise cognitive burden and break-off rates for respondents, perhaps particularly in mobile contexts (Barari et al., 2025).

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

PROOFS

that better reflect respondents' actual latent opinions or by introducing a new source of error.

A closely related concern is the growing prevalence of AI-assisted responding, in which respondents themselves use AI tools when completing online surveys. Early evidence suggests this occurs at substantial rates among experienced nonprobability respondents and poses a serious threat to data integrity (Martherus et al., 2025; Pinzón et al., 2024; Traylor, 2025). For example, LLM-generated responses to open-ended items tend to be more homogeneous and more positive than human-written responses, particularly for sensitive social groups. This potentially distorts downstream representations of public opinion (Zhang et al., 2025).

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

Although no reliable or definitive detectors of AI-assisted responding currently exist, contamination of online survey data is likely to increase as language models improve. Detection strategies based on statistical, linguistic, or behavioral regularities are inherently fragile in adversarial settings, as AI can both learn detection rules and to circumvent them. Even so, researchers have proposed partial tools that may help flag suspicious responses, including metadata signals, paradata anomalies, and semantic diagnostics (Pinzón et al., 2024; Sepulvado, 2025). Open-ended responses can also serve as useful diagnostics of response quality, though overly demanding prompts may inadvertently incentivize AI assistance among respondents (Zhang et al., 2025).

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

These considerations suggest that researchers will require layered approaches combining sampling design, providing survey instructions, monitoring metadata and paradata, and ensuring transparent preregistration of data-quality decisions. More fundamentally, however, the emergence of AI agents capable of autonomously completing surveys unsettles the longstanding assumption that responses can be presumed human. As the boundary

PROOFS

between human and automated responding becomes less reliably detectable, how samples are constructed, monitored, and interpreted becomes central to preserving data quality and inferential validity (Westwood, 2025). Artificial intelligence may reshape survey data collection in productive ways, but its responsible integration depends on deliberate design choices and sustained human oversight.

3.2.3 Post-data Collection

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

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

Performance, however, depends on implementation choices. Classification accuracy varies with prompt design, category definitions, and model configuration (Chae and Davidson 2025), and domain-specific biases can emerge, such as lower accuracy for abortion-related responses expressing pro-life positions. Errors may also vary systematically across respondents and contexts. Ashwin et al. (2025) show that, in interviews with Rohingya refugees, LLMs

PROOFS

overpredict sparse codes and produce false positives correlated with respondent characteristics, such as gender, education, and refugee status. Multilingual and culturally complex data require additional adaptation and careful attention to rare categories (Singh and Kumar, 2025; von der Heyde et al., 2025).

A conceptually distinct set of post-data- collection applications involves using AI to impute missing survey data due to item nonresponse, attrition, or skipped questions. Unlike the coding and analysis tools discussed earlier, which operate on observed responses, imputation implicitly treats model-generated values as estimates of what respondents would have said – a substantively stronger claim whose assumptions are correspondingly harder to evaluate. It also raises an immediate question about the inferential target: Should imputed values approximate the survey answer a respondent would have given, complete with the satisficing, framing effects, and social desirability bias that characterize observed data? Or should they attempt to recover the latent opinion those responses imperfectly reflect? The two targets require different models, different training data, and different standards of validation, yet the distinction is rarely made explicit in practice. Imputation differs from the use of silicon respondents discussed further next because imputation estimates responses conditional on observed human data rather than generating them de novo. Kim and Lee (2026) identify three analytically distinct imputation problems: (1) conventional imputation to generate responses for asked but unanswered questions; (2) retrodiction, or the reconstruction of responses to past questions based on current responses; and (3) unasked opinion prediction, which attempts to infer attitudes on never-asked questions. They show that fine-tuned LLMs trained on representative survey data can perform well for the first two tasks, reconstructing known opinion shifts while maintaining stability during exogenous shocks. But performance declines when predicting unasked opinion and varies considerably across demographic groups.

As with all imputation, using AI to impute survey responses raises important ethical and methodological concerns, particularly when imputing sensitive attitudes or behaviors. Unlike traditional methods, however, the effects of AI-based imputation approaches are harder to evaluate because their assumptions are difficult to specify and interrogate. Conventional imputation models make functional forms, conditioning variables, and missingness mechanisms explicit, whereas AI-based imputations rely on high-dimensional representations learned from opaque training corpora. This opacity increases the risk that model-generated responses default to typical patterns, overstate precision, mask heterogeneity, and blur the distinction between measured responses and model-based inference. Scholars must therefore distinguish carefully between observed data and AI-derived estimates.

Artificial intelligence's post-data- collection uses range from efficiency-enhancing tools for coding and analysis to more consequential forms of inference that reshape how missing data are handled. We currently lack a systematic assessment of how these applications affect Total Survey Error (but

PROOFS

see Rothschild et al., 2026), underscoring the need for caution and clearer evaluative standards as AI becomes embedded in post-survey workflows. The most extreme form of this progression in which AI-generated data replace that of human respondents entirely is the subject to which we now turn.

3.3 SILICON RESPONDENTS

Perhaps the most transformational, and controversial, way in which AI reshapes the conceptualization and measurement of public opinion is through the creation of silicon respondents (Argyle et al., 2023; Horton et al., 2026; Park et al., 2026; Wang et al., 2025). Silicon respondents are often generated by prompting LLMs to adopt a persona defined by demographic, attitudinal, or contextual information – and answer survey questions as that persona to approximate how members of the adopted target population might respond. Unlike classical imputation or synthetic data techniques grounded in explicit probabilistic models such as multiple imputation by chained equations or predictive mean matching (Drechsler and Haensch, 2024; Latner et al., 2024), AI-generated responses do not emerge from transparent statistical assumptions about sampling or missingness. Instead, they reflect an opaque combination of prompts, patterns learned from massive but not necessarily representative training corpora, and post-training alignment procedures.

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

PROOFS

confronted explicitly in ways that conventional survey research has often been able to leave implicit.

Independent of this conceptual ambiguity is the difficulty of establishing clear evaluative criteria. Conclusions about the ability of silicon respondents to mimic human responses depend heavily on the properties being assessed. Concerns are often framed in terms of algorithmic bias arising from training data (Bender and Friedman, 2018), post-training alignment (Lyman et al., 2025), or prompt design (Atreja et al., 2024; see also Gallegos et al., 2024; Hu et al., 2025; Westwood et al., 2025), but the validity of silicon respondents depends on the aggregate effects of each. The central question is not whether a model exhibits bias in how it responds to a particular prompt, but whether it can adopt personas corresponding to real populations of interest and generate responses that accurately reflect the joint distribution of human opinions both within and across questions. Scholars have proposed a range of criteria for evaluating silicon respondents (Röttger et al., 2024). Early work emphasized overall correspondence between human and silicon responses (Argyle et al., 2023), while subsequent studies focused on variance and distributional properties (Bisbee et al., 2024) and subgroup-level correspondence within populations (Santurkar, 2023; Kim and Lee, 2026). Different approaches prioritize different evaluative targets and reach divergent conclusions about the usefulness and reliability of silicon respondents.

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

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

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

PROOFS

benchmarks does not imply fidelity going forward, particularly as events, contexts, and opinions change.

A second limitation concerns what algorithmic fidelity is actually measuring. The standard benchmark treats correspondence with observed human survey responses as the validation criterion. But observed sample responses are not the same as population-level opinion, and the gap between them matters. With human surveys, the sources of nonrepresentativeness are at least partially observable, and researchers can in principle adjust for them through weighting and post-stratification. With silicon respondents, the effective sampling frame is the LLM's training data, which is opaque: It is not possible to know which populations are over- or underrepresented in ways that would allow principled correction. Matching a particular human sample, therefore, does not establish that silicon respondents recover population-level opinion. The problem is compounded if the goal is latent opinion rather than observed survey responses. If what we care about is what people actually believe, independent of the artifacts introduced by the survey process itself, then fidelity to survey responses embeds the very distortions, including satisficing, framing effects, and social desirability bias, that the concept of latent opinion is defined against. In short, algorithmic fidelity is most defensible precisely when it is least needed, when high-quality representative human survey data already exist, and least defensible where silicon respondents are most appealing, when such data are unavailable, unrepresentative, or potentially distorted. Alternative benchmarks assessing correspondence with population-level behavioral outcomes, such as turnout, election results, or political donations, may be more appropriate in some contexts (e.g., Cerina and Duch, 2025), though these too carry assumptions about what counts as ground truth.

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

In practice, the use of silicon respondents frequently proceeds without resolving this ambiguity, leaving uncertain both what is being measured and whether the use of silicon respondents reflects institutional or economic incentives, rather than a coherent position on what public opinion is and how it ought to be measured. Given this unresolved ambiguity, it is useful to consider

PROOFS

the spectrum of ways in which silicon respondents may be used, ordered roughly by the importance of resolving the underlying conceptual question.

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

Beyond this diagnostic role, silicon respondents may also have instrumental value for study design and power analysis. They can be used to explore hypothetical relationships much as agent-based models do – assessing whether planned analyses appear able to detect anticipated effect sizes or determining required sample sizes given plausible effect distributions (e.g., Sarstedt et al., 2024; Dillon et al., 2023; Ashokkumar et al., 2026; Lippert et al., 2024). In these applications, silicon respondents are not obviously inferior to existing practices, but caution is warranted because of possible silicon demand effects, in which AI systems infer experimental hypotheses and generate responses that artificially confirm them (Westwood, 2025).

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

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

The possibility of model collapse exposes a broader limitation of prevailing validation strategies. A researcher may save resources or improve performance in a single application, yet in the aggregate such choices can erode the

PROOFS

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

An intermediate, hybrid approach combines silicon data with human survey data in an effort to reduce bias under conditions of extreme nonresponse or selective participation (e.g., Berinsky, 2004; Kennedy et al., 2018; Bailey, 2024). Rather than replacing human respondents altogether, hybrid designs use silicon respondents to approximate the views of individuals or groups who exist in the population but do not participate in surveys (Cerina and Duch, 2025; Cerina and Duch, 2025). In principle, such augmentation may improve inference if human-only samples are increasingly unrepresentative.

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

Although the use of silicon responses may reduce costs and expand access, particularly for resource-poor teams or hard-to-survey populations, neither advantage ensures validity nor reliability. Proprietary frontier models may offer superior performance at the expense of transparency and reproducibility relative to open-weight alternatives, and performance may vary greatly based on pretraining, prompting, and post-model calibration. As a result, nominally similar silicon respondents may reflect meaningfully different assumptions and behaviors across models and implementations, making it difficult to determine whether observed differences arise from substantive features of public opinion, from contingent modeling choices, or from changes in the model systems

PROOFS

themselves. Using silicon respondents therefore requires reasoned arguments about the meaning of public opinion and the risks researchers are prepared to accept in the absence of a clear or stable ground truth.

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

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

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

3.4 TRANSPARENCY, DISCLOSURE, AND REPLICATION

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

PROOFS

The integration of AI further complicates traditional notions of replication and reproducibility (see Chapter 10). As with survey research itself, replication of AI-integrated tools or data cannot mean simply repeating the same procedure and obtaining identical results because models are frequently retrained and altered without notice, and stochastic generation introduces variation across runs due to random seeds, parameter settings, or minor prompt changes beyond the researcher's control.

These characteristics raise a deeper conceptual challenge: What does replication mean when both the tool and its internal states evolve over time? Perfect replication is extremely difficult when studying digitally mediated phenomena, due to the problem of temporal validity (Munger, 2023). As more politically relevant behavior, including the measurement of public opinion, incorporates generally nondeterministic AI throughout the research process, the problem of temporal validity worsens and the notion of replication shifts from a standard of duplication toward documenting conditions of use and assessing robustness across plausible configurations. Taken together these issues shift the focus toward reproducibility of the process rather than replication of the exact results.

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

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

PROOFS

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

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

We recognize that the increasing integration of AI will make it difficult to document every use of AI across all survey research workflows. Nevertheless, it is essential that, at a minimum, researchers document how their use of AI could plausibly affect the data, the estimand, and the resulting conclusions. The goal is to provide sufficient transparency to allow others to trace, understand, and, when possible, replicate or at least partially reproduce the processes used to generate AI-assisted responses or analyses so that others may evaluate how AI shapes knowledge claims being made. At the same time, researchers may be unaware of how AI systems shape outputs (Cheng et al., 2023), reinforcing the need for transparency standards that support independent evaluation.

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

The second form is *conceptual or interpretive transparency*, which concerns how researchers conceptualize and justify their use of AI. Beyond purely efficiency-enhancing applications, such as assistance with formatting, code writing, or minor editing, the use of AI is not a neutral technical detail. Instead, it constitutes a theoretical and methodological commitment that warrants

PROOFS

justification proportional to its potential inferential consequences. As we have argued, when AI tools shape what data are collected, how responses are generated or interpreted, or which quantities are ultimately estimated, they embed assumptions about what public opinion is, how it can be observed, and what sources of error are acceptable. At a minimum, researchers should be explicit about which of the three targets their use of AI is aimed at: observed sample responses, latent respondent opinion, or population-level inference. This choice determines what fidelity means, what benchmarks are appropriate, and what claims the resulting data can legitimately support. These assumptions also condition the relationship between observed data and the substantive claims made about publics, attitudes, and behavior. Researchers should clearly describe where AI substitutes for or reshapes human input, what assumptions any substitution entails, and how it could plausibly affect substantive conclusions. In this sense, incorporating AI into public opinion research more closely resembles adopting a modeling framework than adopting a new software tool and, therefore, demands explicit and transparent justification of its conceptual foundations, as well as the statistical and epistemological risks it entails and its fitness for use.

3.5 OUTSTANDING QUESTIONS

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

- As AI increasingly shapes, mediates, or proxies for human beliefs, does this alter what we mean by public opinion, or merely how we observe it?
- How should we interpret LLM-based summaries of human expression? On the one hand, they expand the information that can be used to characterize human opinion beyond survey responses, but, on the other hand, it is unclear how those expressions relate to human populations of interest.
- Given that LLMs overrepresent text-rich groups and high-resource languages, do AI-generated summaries or silicon respondents amplify the voices of the already-empowered? If so, how should researchers account for these distortions?
- As work continues to evaluate the performance of silicon respondents as a proxy for human opinion, what are the implications of the fact that available human survey benchmarks are themselves historical, unstable, and subject to survey error?
- What norms should govern the use and disclosure of AI in opinion measurement and political communication? How can political science maintain scientific norms of disclosure and replication when outputs result from opaque and changing processes?

PROOFS

- Normatively, what happens to democratic legitimacy when the line between measuring and manufacturing opinion collapses? Particularly as silicon respondents and AI-based persuasive systems are deployed at scale?
- Does the field need a unified framework that is analogous to the Total Survey Error paradigm to characterize the myriad ways that AI may affect the production, expression, and measurement of public opinion? AI introduces new forms of possibly correlated error at every stage of the opinion lifecycle – from opinion formation to generation and expression to measurement.

3.6 CONCLUSIONS AND IMPLICATIONS

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

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

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

PROOFS

The paradox is that as AI becomes easier and more capable, our responsibility as scholars only increases. Greater computational power does not absolve researchers from grappling with theory, uncertainty, and normative judgment; it intensifies those obligations. Public opinion research has long been justified not only as a scientific enterprise, but as a democratic effort to make the preferences, beliefs, and concerns of citizens visible and consequential. If AI systems increasingly mediate, approximate, or replace those voices, the central question is not whether the technology works but whether its use preserves the epistemic and democratic commitments that have motivated the study of public opinion.

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

PROOFS

BIBLIOGRAPHY

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

Adhikari, D. M., Hartland, A., Weber, I., & Cannanure, V. K. (2025). Exploring LLMs for Automated Generation and Adaptation of Questionnaires. *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, 1–23. https://doi.org/10.1145/3719160.3736606

Allamong, M. B., Jeong, J., & Kellstedt, P. M. (2025). Spelling correction with large language models to reduce measurement error in open-ended survey responses. *Research & Politics*, 12(1). https://doi.org/10.1177/20531680241311510

Alshaabi, T., Adams, J. L., Arnold, M. V., et al. (2021). Storywrangler: A Massive Exploratorium for Sociolinguistic, Cultural, Socioeconomic, and Political Timelines Using Twitter. *Science Advances*, 29(7). https://doi.org/10.1126/sciadv.abe6534

Amirova, A., Fteropoulli, T., Ahmed, N., Cowie, M. R., & Leibo, J. Z. (2024). Framework-Based Qualitative Analysis of Free Responses of Large Language Models: Algorithmic Fidelity. *PLOS ONE*, 19(3), e0300024. https://doi.org/10.1371/journal.pone.0300024

Anthis, J. R., Liu, R., Richardson, S. M., et al. (2025). Position: LLM Social Simulations Are a Promising Research Method. In *Forty-second International Conference on Machine Learning Position Paper Track*. https://proceedings.mlr.press/v267/anthis25a.html

Argyle, L. P., Busby, E. C., Fulda, N., et al. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3), 337–351. https://doi.org/10.1017/pan.2023.2

Argyle, L. P., Busby, E. C., Gubler, J. R., et al. (2025). Arti-"Fickle" Intelligence: Using LLMs as a Tool for Inference in the Political and Social Sciences. *Nature Computational Science*, 5, 737–744. https://doi.org/10.1038/s43588-025-00843-4

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

Ashokkumar, A., Hewitt, L., Ghezae, I., & Willer, R. (2026). Large Language Models Can Predict the Results of Social Science Experiments. *Nature*. Advance online publication. https://doi.org/10.1038/s41586-026-10742-x

Ashwin, J., Chhabra, A., & Rao, V. (2025). Using Large Language Models for Qualitative Analysis Can Introduce Serious Bias. *Sociological Methods & Research*, 55(3). https://doi.org/10.1177/00491241251338246

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

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

Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2025). Explicitly Unbiased Large Language Models Still Form Biased Associations. *Proceedings of the National Academy of Sciences*, 122(8), e2416228122. https://doi.org/10.1073/pnas.2416228122

Barari, S., Angbazo, J., Wang, N., et al. (2025). AI-Assisted Conversational Interviewing: Effects on Data Quality and Respondent Experience. *arXiv preprint*. arXiv:2504.13908. https://doi.org/10.48550/arXiv.2504.13908

Barberá P, Boydstun AE, Linn S, McMahon R, Nagler J. (2021). Automated Text Classification of News Articles: A Practical Guide. *Political Analysis*. 29(1), 19–42. https://doi.org/10.1017/pan.2020.8

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

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

Behrend, T. S., & Landers, R. N. (2025). Participant Interactions with Artificial Intelligence: Using Large Language Models to Generate Research Materials for Surveys and Experiments. *Journal of Business and Psychology*, 40(6), 1275–1297. https://doi.org/10.1007/s10869-025-10035-6

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

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

Bisbee, J., Clinton, J. D., Larson, J. M, & Lee, D. I. (2026). AI Pandering: Constructing Diverging Political Realities through Conversation. *APSA Preprints*. https://doi.org/10.33774/apsa-2026-ccvw9

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

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

PROOFS

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

Burke, E. (1774). Speech at Mr. Burke's Arrival at Bristol, and at the Conclusion of the Poll. In F. Laurence & W. King (Eds.), *The Works of the Right Honourable Edmund Burke* (Vol. III). London: F. & C. Rivington.

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

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

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

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

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

Chae, Y., & Davidson, T. (2025). Large Language Models for Text Classification: From Zero-Shot Learning to Instruction-Tuning. *Sociological Methods & Research*, 55(2), 501–567. https://doi.org/10.1177/00491241251325243

Cheng, M., Piccardi, T., & Yang, D. (2023). CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations. arXiv preprint arXiv:2310.11501. https://arxiv.org/abs/2310.11501

Chintakunta, S. S., Nascimento, N., & Guimaraes, E. (2025). Large Language Models in the Data Science Lifecycle: A Systematic Mapping Study. arXiv preprint arXiv:2508.11698. https://arxiv.org/abs/2508.11698

Clinton, J. D., Agiesta, J., Brenan, M., et al. (2021). Task Force on 2020 Pre-Election Polling: An Evaluation of the 2020 General Election Polls. American Association of Public Opinion Research. https://tinyurl.com/53hu8ssc

Converse, P. E. (1964). The Nature of Belief Systems in Mass Publics. In Apter, D.E. (Ed.), *Ideology and Discontent* (pp. 206–261). New York: The Free Press.

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably Reducing Conspiracy Beliefs through Dialogues with AI. *Science*, 385(6714), eadq1814. https://doi.org/10.1126/science.adq1814

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

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

PROOFS

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

Dillman, D. A. (2000). *Mail and Internet Surveys: The Tailored Design Method* (2nd ed.). New York: Wiley.

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

Duch, R., Batistoni, T., Beretta, M., & Low, R. (2026, June 28). Synthetic Participants for Real Interventions: Augmenting Behavioral RCTs with Fine-Tuned LLMs. https://doi.org/10.17605/OSF.IO/68JFQ

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

Freelon, D. (2018). Computational Research in the Post-API Age. *Political Communication*, 35(4), 665–668. https://doi.org/10.1080/10584609.2018.1477506

Fuchs, A., Haensch, A.-C., & Weber, W. (2026). AI for Survey Design: Generating and Evaluating Survey Questions with Large Language Models. *SocArXiv*. https://sciety.org/articles/activity/10.31235/osf.io/fzn7t_v1

Gallegos, I. O., Rossi, R. A., Barrow, J., et al. (2024). Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3), 1097–1179. https://doi.org/10.1162/coli_a_00524

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

Ghosh, S., & Caliskan, A. (2023). ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 901–912. https://doi.org/10.1145/3600211.3604672

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

Groves, R. M., Fowler, F. J., Couper, M. P., et al. (2009). *Survey Methodology*. (2nd ed.). Hoboken, NJ: Wiley.

Haavisto, O., & Welsch, R. (2024). Questionnaires for Everyone: Streamlining Cross-Cultural Questionnaire Adaptation with GPT-Based Translation Quality Evaluation. *arXiv preprint.* arXiv:2407.20608. https://arxiv.org/abs/2407.20608

Hackenburg, K., & Margetts, H. (2024). Evaluating the Persuasive Influence of Political Microtargeting with Large Language Models. *Proceedings of the National Academy of Sciences*, 121, e2403116121o. https://doi.org/10.1073/pnas.2403116121

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

Hernandez, I., & Nie, W. (2023). The AI-IP: Minimizing the Guesswork of Personality Scale Item Development through Artificial Intelligence. *Personnel Psychology*, 76(4), 1011–1035. https://doi.org/10.1111/peps.12543

Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). AI Generates Covertly Racist Decisions about People Based on Their Dialect. *Nature*, 633, 147–154. https://doi.org/10.1038/s41586-024-07856-5

Horton, J. J. Filippas, A., & Manning, B.S. (2026). Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? [NBER Working Paper 31122]. https://doi.org/10.3386/w31122

PROOFS

Hu, T., Kyrychenko, Y., Rathje, S., et al. (2025). Generative Language Models Exhibit Social Identity Biases. *Nature Computational Science*, 5, 65–75. https://doi.org/10.1038/s43588-024-00741-1

Huang, Y., Wu, R., He, J., & Xiang, Y. (2024). Evaluating ChatGPT-4.0's Data Analytic Proficiency in Epidemiological Studies: A Comparative Analysis with SAS, SPSS, and R. *Journal of Global Health*, 14, 04070. https://doi.org/10.7189/jogh.14.04070

Ho, X., Wu, Y.-A., Kumar, S., et al. (2025). Format Matters: The Robustness of Multimodal LLMs in Reviewing Evidence from Tables and Charts. *arXiv preprint*. arXiv:2511.10075. https://doi.org/10.48550/arXiv.2511.10075

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

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

Jamieson, K. H., Lupia, A., Amaya, A., et al. (2023). Protecting the Integrity of Survey Research. *PNAS Nexus*, 2(3), pgad049. https://doi.org/10.1093/pnasnexus/pgad049

Jansen, P., Tafjord, Ø., Radensky, M., et al. (2025). CodeScientist: End-to-End Semi-Automated Scientific Discovery with Code-Based Experimentation. In *Findings of the Association for Computational Linguistics: ACL 2025*, 13370–13467. https://doi.org/10.18653/v1/2025.findings-acl.692

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

Kennedy, C., Blumenthal, M., Clement S., et al. (2018). An Evaluation of the 2016 Election Polls in the United States: AAPOR Task Force Report. *Public Opinion Quarterly*, 82(1), 1–33. https://doi.org/10.1093/poq/nfx047

Key, V. O. (1961). Public Opinion and American Democracy. New York: Alfred A. Knopf.

Kim, J., & Lee, B. (2026). AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction. *arXiv preprint*. arXiv:2305.09620v4. https://arxiv.org/abs/2305.09620v4

Krumpal, I. (2013). Determinants of Social Desirability Bias in Sensitive Surveys: A Literature Review. *Quality & Quantity*, 47(4), 2025–2047. https://doi.org/10.1007/s11135-011-9640-9

Lang, M. M., & Eskenazi, S. (2025). Telephone Surveys Meet Conversational AI: Evaluating a LLM-Based Telephone Survey System at Scale. *arXiv preprint*. arXiv:2502.20140. https://doi.org/10.48550/arXiv.2502.20140

Latner, J., Neunhoeffer, M., & Drechsler, J. (2024). Generating Synthetic Data Is Complicated: Know Your Data and Know Your Generator. In J. Domingo-Ferrer & M. Önen (Eds.), *Privacy in Statistical Databases 2024*, 115–128. https://doi.org/10.1007/978-3-031-69651-0_8

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

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

PROOFS

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

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

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

Lin, H., Czarnek, G., Lewis, B., et al. (2025). Persuading Voters Using Human–Artificial Intelligence Dialogues. *Nature*, 648(8093), 394–401. https://doi.org/10.1038/s41586-025-09771-9

Lippert, S., Dreber, A., Johannesson, M., et al. (2024). Can Large Language Models Help Predict Results from a Complex Behavioural Science Study? *Royal Society Open Science*, 11(9), 240682. https://doi.org/10.1098/rsos.240682

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

Lyman, A., Hepner, B., Argyle, L. P., et al. (2025). Balancing Large Language Model Alignment and Algorithmic Fidelity in Social Science Research. *Sociological Methods & Research*, 54(3), 1110–1155. https://doi.org/10.1177/00491241251342008

Maiorino, A., Padgett, Z., Wang, C., Yakubovskiy, M., & Jiang, P. (2023). Application and Evaluation of Large Language Models for the Generation of Survey Questions. *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 5244–5245. https://doi.org/10.1145/3583780.3615506

Martherus, J., Podkul, A., Cook, E., & Liebowitz, R. (2025). How to Detect AI-Assisted Interviews in Online Surveys. *Survey Practice*, 18(1), 1–10. https://doi.org/10.29115/SP-2025-0016

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

Metheney, E. A., & Yehle, L. (2024). Exploring the Potential Role of Generative AI in the TRAPD Procedure for Survey Translation. *arXiv preprint*. arXiv:2411.14472. https://doi.org/10.48550/arXiv.2411.14472

Munger, K. (2020). All the News that's Fit to Click: The Economics of Clickbait Media. *Political Communication*, 37(3), 376–397. https://doi.org/10.1080/10584609.2019.1687626

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

Munger, K. (2024). *The YouTube Apparatus (Elements in Politics and Communication)*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781009359795

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

Olivos, F., & Liu, M. (2024). ChatGPTest: Opportunities and Cautionary Tales of Utilizing AI for Questionnaire Pretesting. *arXiv preprint*. arXiv:2405.06329. https://doi.org/10.48550/arXiv.2405.06329

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

PROOF

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

Palmer, A., Smith, N. A., & Spirling, A. (2024). Using Proprietary Language Models in Academic Research Requires Explicit Justification. *Nature Computational Science*, 4(1), 2–3. https://doi.org/10.1038/s43588-023-00585-1

Park, J. S., Zou, C. Q., Kamphorst, J., et al. (2026). LLM Agents Grounded in Self-Reports Enable General-Purpose Simulation of Individuals. *arXiv preprint*. arXiv:2411.10109. https://arxiv.org/abs/2411.10109

Pinzón, N., Koundinya, V., Galt, R. E., et al. (2024). AI-Powered Fraud and the Erosion of Online Survey Integrity: An Analysis of 31 Fraud Detection Strategies. *Frontiers in Research Metrics and Analytics*, 9, 1432774. https://doi.org/10.3389/frma.2024.1432774

Reja, U., Lozar Manfreda, K., Hlebec, V., & Vehovar, V. (2003). Open-Ended vs. Close-Ended Questions in Web Questionnaires. *Developments in Applied Statistics*, 19, 159–177. http://mrvar.fdv.uni-lj.si/pub/mz/mz19/reja.pdf

Rothschild, D. M., Buskirk, T. D., Eckman, S., et al. (2025). Successfully Navigating the Disruption AI Will Bring to Survey Research. *The Survey Statistician*, 92, 30–44.

Rothschild, D., Marlar, J., Amaya, A., et al. (2026). Responsible AI Integration in Survey Research. *American Association for Public Opinion Research*. https://tinyurl.com/37rhkvvh

Röttger, P., Hofmann, V., Pyatkin, V., et al. (2024). Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 15295–15311). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2024.acl-long.816

Rytting, C. M., Sorensen, T., Argyle, L., et al. (2023). Towards Coding Social Science Datasets with Language Models. *arXiv preprint*. arXiv:2306.02177. https://doi.org/10.48550/arXiv.2306.02177

Santurkar, S., Durmus, E., Ladhak, F., et al. (2023). Whose Opinions Do Language Models Reflect? In *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202, pp. 29971–30004). PMLR. https://proceedings.mlr.press/v202/santurkar23a.html

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

Schroeder, D. T., Cha, M., Baronchelli, A., et al. (2026). How Malicious AI Swarms Can Threaten Democracy: The Fusion of Agentic AI and LLMs Marks a New Frontier in Information Warfare. *Science*, 391(6783), 354–357. https://doi.org/10.1126/science.adz1697

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

Sepulvado, B. (2025, December). *Detecting AI Responses in Survey Data: NORC's Next Leap for Data Quality*. NORC at the University of Chicago. https://tinyurl.com/3y4r679k

PROOFS

Settle, J. E. (2018). *Frenemies: How Social Media Polarizes America*. Cambridge: Cambridge University Press.

Shumailov, I., Shumaylov, Z., Zhao, Y., et al. (2024). AI Models Collapse When Trained on Recursively Generated Data. *Nature*, 631(8022), 755–759. https://doi.org/10.1038/s41586-024-07566-y

Singh, T., & Kumar, H. (2025). AI in Qualitative Research: Using Large Language Models to Code Survey Responses in Native Languages. *IFPRI Blog*, July 10. https://tinyurl.com/2paa2pjt

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

Spirling, A. (2023). Why Open-Source Generative AI Models Are an Ethical Way Forward for Science. *Nature*, 616(7957), 413. https://doi.org/10.1038/d41586-023-01295-4

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural Bias and Cultural Alignment of Large Language Models. *PNAS Nexus*, 3(9), pgae346. https://doi.org/10.1093/pnasnexus/pgae346

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

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

Traylor, F. (2025). The Threat of AI Chatbot Responses to Crowdsourced Open-Ended Survey Questions. *Energy Research & Social Science*, 119, 103857. https://doi.org/10.1016/j.erss.2024.103857

Velez Y. R. (2025). Crowdsourced adaptive surveys. *Political Analysis* 33(4), 284–297. https://doi.org/10.1017/pan.2024.34

Velez Y. R., & Liu P. (2025). Confronting core issues: A critical assessment of attitude polarization using tailored experiments. *American Political Science Review* 119(2), 1036–1053. https://doi.org/10.1017/S0003055424000819

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

Waight, H., Yang, E., Yuan, Y., et al. (2026). State Media Control Influences Large Language Models. *Nature*, 655, 685–693. https://doi.org/10.1038/s41586-026-10506-7

Wang, A., Morgenstern, J., & Dickerson, J. P. (2025). Large Language Models that Replace Human Participants Can Harmfully Misportray and Flatten Identity Groups. *Nature Machine Intelligence*, 7(3), 400–411. https://doi.org/10.1038/s42256-025-00986-z

Wang, Q., Wu, J., Jiang, Z., et al. (2025). LLM-based Human Simulations Have Not Yet Been Reliable. *arXiv preprint*. arXiv:2501.08579. https://doi.org/10.48550/arXiv.2501.08579

West, B. T., & Blom, A. G. (2017). Explaining Interviewer Effects: A Research Synthesis. *Journal of Survey Statistics and Methodology*, 5(2), 175–211. https://doi.org/10.1093/jssam/smw024

PROOFS

Westwood, S.J. (2025). The potential existential threat of large language models to online survey research. *Proceedings of the National Academy of Sciences*, 122(47), e2518075122. https://doi.org/10.1073/pnas.2518075122

Westwood, S. J., Grimmer, J., & Hall, A. B. (2025). Measuring Perceived Slant in Large Language Models Through User Evaluations. Stanford Graduate School of Business Working Paper No. 4262. https://tinyurl.com/37wpfxnt

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

Xiao, Z., Zhou, M. X., Liao, Q. V., et al. (2020). Tell Me About Yourself: Using an AI-Powered Chatbot to Conduct Conversational Surveys with Open-Ended Questions. *ACM Transactions on Computer-Human Interaction*, 27(3), 1–37. https://doi.org/10.1145/3381804

Yun, H. S., Arjmand, M., Sherlock, P., et al. (2024). Keeping Users Engaged During Repeated Interviews by a Virtual Agent: Using Large Language Models to Reliably Diversify Questions. *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*. https://doi.org/10.1145/3652988.3673929

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

Zhang, S., Xu, J., & Alvero, A. (2025). Generative AI Meets Open-Ended Survey Responses: Research Participant Use of AI and Homogenization. *Sociological Methods & Research*, 54(3), 1197–1242. https://doi.org/10.1177/00491241251327130

PROOFS

4

AI, the Public Sector, and Policymaking

Baobao Zhang, Diane Coyle, Jae Yeon Kim, Johannes Himmelreich, and Mila Gascó-Hernandez

PROOFS

4.1 INTRODUCTION

The city council of Porto Alegre, Brazil, unanimously passed a bill in October 2023, exempting residents from having to pay to replace stolen water meters. On its face, the bill was routine. What made it historic was revealed only after passage: the bill had been written entirely by ChatGPT – the first known law to be entirely written by artificial intelligence (AI). Councilman Ramiro Rosário had entered a forty-nine-word prompt and received a draft proposal within seconds. He deliberately concealed this fact from his fellow council members because he believed they would not vote for an AI-generated bill (Jeantet and Savarese 2023). Although Rosário said he believed that using AI to write legislation would increase legislative efficacy, Porto Alegre City Council President Hamilton Sossmeier disagreed: "[I]t's a dangerous precedent. It's complicated. Other laws that are more complex may follow, generating another type of impact" (Fleck 2024).

This short example highlights opportunities and challenges of AI in the public sector. Artificial intelligence holds out the promise to improve efficiency,

AI use disclosure: The authors used Gemini 2.5 Pro to generate the detailed outline for the paper. The prompt used to generate the detailed outline included the content provided by the co-authors above and a preliminary outline developed by Baobao Zhang and Johannes Himmelreich. Baobao Zhang further revised the outline independently, taking into account the co-authors' comments. The authors used Claude Opus 4.5 and Gemini 3 for literature review and cleaning up the citations/bibliography. The authors checked the citations manually and used Grammarly to correct spelling and grammatical errors.

Conflict of interest disclosure: Diane Coyle is an academic adviser on digital markets to the United Kingdom Competition and Markets Authority and is a member of Google's advisory group on AI and the economy. Co-author Jae Yeon Kim's spouse works for Amazon.

Tucker_9781047774734_C004.indd 109 15-09-2026 08:55:42

textual clarity, and expertise in policymaking (Sanders and Schneier 2025). Artificial intelligence can help write complex legislation that would otherwise require vast human resources. Artificial intelligence can identify compromise statements among citizens and representatives who disagree. And AI could identify relevant precedent legislation or identify shortcomings in legislative drafts. Such opportunities, however, also present challenges. Artificial intelligence development and use are largely nontransparent – for technical and organizational reasons. Technically, the AI models are typically inscrutable, and their decisions are hard to explain (Papadakis et al. 2024). Organizationally, the public sector is increasingly relying on technical systems and services developed by private vendors. This chapter structures the benefits and identifies the challenges that AI raises for the public sector.

Governments around the world have been using algorithms and machine learning for nearly a decade to automate service provision and internal management. But the increasing capability and deployment of generative AI means the technology can become a potential author of policy itself (Lessig 1999; Bar-Siman-Tov 2025). The Porto Alegre case illustrates how governments adopt AI through informal experimentation, rather than through thoughtful institutional design, thus bypassing public disclosure and democratic deliberation that legitimize public decisions.

The integration of AI into government is a transformation that probes public service values, challenges state capacity, and opens accountability gaps. First, deploying AI in the service of the public interest creates tensions with the private sector, which develops these tools primarily for optimization and profit. Information and power asymmetries between technology companies and governments render it increasingly difficult to implement AI aligned with the public good. Second, decades of outsourcing to private tech companies have hollowed out the state's technical capacity, leaving governments dependent on vendors they cannot effectively oversee. Third, when decisions are delegated to AI systems, existing accountability mechanisms – vertical, horizontal, and diagonal – break down, making it difficult to redress harms.

The chapter proceeds as follows. Section 4.2 presents a taxonomy of governmental AI uses. It illustrates and structures the benefits of AI by mapping thirteen categories of AI activity onto three core governance functions (i.e., policymaking, public service delivery, and internal management), with examples from local, regional, and national governments worldwide. Section 4.3 examines three interconnected challenges: aligning public service values with algorithmic optimization, rebuilding state capacity in an era when governments depend on private vendors, and developing accountability mechanisms for the algorithmic state. Section 4.4 proposes a research agenda addressing these challenges, emphasizing the foundational need for transparency, and outlining key questions for political scientists and public administration scholars.

PROOFS

4.2 GOVERNMENTS' USE OF AI

Artificial intelligence use by governments has proliferated in recent years. Countries in the Global North, in particular, have robustly expanded digital government, R&D, and data infrastructure. According to a 2024 report from the Organisation for Economic Co-operation and Development (OECD), 70 percent of its member countries use AI systems for internal government processes, 67 percent for public service design and delivery, and 33 percent for policymaking (OECD 2024a). By analyzing 200 real-world AI use cases, the report found that 57 percent of deployed AI systems support automating, streamlining, or tailoring public services, while 45 percent enhance decision-making and forecasting capabilities. In the United States alone, the Government Accountability Office (GAO), often called the US congressional watchdog, reports that, across eleven major federal agencies, AI use cases nearly doubled – from 571 in 2023 to 1,110 in 2024. Generative AI applications increased nearly ninefold – from 32 to 282 implementations – during this same period (US Government Accountability Office 2025).

To structure how, where, and why governments use AI, this section presents a taxonomy of AI use in government by synthesizing key elements from existing frameworks. Table 4.A showcases this taxonomy and provides examples of how local, state/regional, and national governments use AI, including proof-of-concept, pilot programs, and legacy programs. Finally, this section offers an overview of how AI is currently being used by governments, highlighting salient current trends.

The taxonomy presented in this chapter combines the National Institute of Standards and Technology (NIST) *AI Use Taxonomy* (Theofanos, Choong, and Jensen 2024) with Van Noordt and Misuraca's (2022) three broad governance functions. The NIST *AI Use Taxonomy* is the most up-to-date and comprehensive classification system for AI use at the time of writing. The NIST taxonomy includes sixteen AI use activities that are independent of AI techniques and domains (e.g., decision-making can deploy a traditional machine learning risk scoring algorithm or a large language model (LLM)). We consolidate some of the uses based on their functional similarity. Unlike previous AI use taxonomies, this taxonomy includes use activities in which generative AI excels, such as content creation and synthesis. Crucially, this taxonomy is general enough to encompass future use cases in the public sector. The taxonomy maps these general AI use activities to three governance functions: policymaking, public service delivery, and internal management.

- *Policymaking* includes activities such as policy analysis, drafting legislation, rulemaking, and communicating with the public about policy.
- *Public service delivery* includes providing citizen-facing services and administering benefits.
- *Internal management* includes back-office operations, human resources, organizational management, and procurement of products or services.

PROOFS

These three functions are broad and can span across multiple branches of government. For example, policymaking can occur in the legislative branch, through the passage of broad laws, or by executive agencies, through administrative rule-making.

This taxonomy of government AI use (Table 4.A) highlights ongoing trends. Whereas prediction and decision support have a long history in data science and AI, some newer AI use cases remain exploratory. For instance, local governments are experimenting with content creation and content synthesis, such as drafting policy documents, summarizing meeting notes, and synthesizing existing documents that are made possible by LLMs in the policymaking process. Other uses of generative AI, such as informing optimal economic policy, may have been some early implementations, but we lack their official documentation and strong evaluations of their impact, or they have not scaled up. We indicate documented prototypical applications as *proof-of-concept* in Table 4.A.

The cells in the taxonomy are not mutually exclusive. For instance, most AI-based image analysis serves the purpose of detection (e.g., identifying potential suspects from a passport photo database) or monitoring (e.g., continuous satellite image monitoring of illegal deforestation). Nevertheless, defining AI activities too broadly might render distinctions meaningless. Finally, general-purpose AI, such as LLMs, which some governments are increasingly adopting, can perform a wider range of activities (e.g., content generation, content synthesis, decision support, discovery, and information retrieval) than traditional machine learning models designed for specific purposes.

4.3 CORE CHALLENGES

Given this overview of how, where, and why governments are using AI, Sections 4.3.1–4.3.3 examine three interconnected governance challenges that these technologies introduce.

We identify three challenges as particularly pressing: a conflict between *public service values and algorithmic optimization*, a growing *information and power asymmetry*, and, as a result, *accountability gaps* when governmental decisions are delegated to opaque, proprietary systems. Successfully navigating these challenges requires clear policy guidelines, stronger state capacity, and robust accountability frameworks.

4.3.1 Challenge 1: AI Conflicts with Public Service Values

The development and deployment of AI involve key design decisions, such as those regarding outcomes, model form, data, objective function, and interaction design (Ludwig and Mullainathan 2021). These design decisions involve trade-offs, require judgment and, hence, are made, if only implicitly, based

PROOFS

on values.[1] These key design decisions encode, and often optimize, some values over others.

Unlike the private sector's metrics of profit and efficiency, the public sector is guided by a pursuit of the public interest, which entails a commitment to a multitude of *public values* such as participation, due process, and cost efficiency (Bozeman 2017; Nabatchi 2018).[2] The use of AI in the public sector, thus, needs to contend with the particular challenges posed by public values.

(1) *Multidimensionality*. Public service has normative and multidimensional objectives. Some of those objectives play a more prominent role in some contexts than others. For drinking water or mail delivery, universal service coverage trumps efficiency: All households ought to have access to water and mail, even if not all service areas are profitable. The US Postal Service (USPS) serves areas of the country where profit-driven organizations would never go (Yager 2025). Military logistics and operations aim for effectiveness and safety – at the expense of efficiency (several layers of redundancy can be costly but, hopefully, will not be needed). But when 200 miles of highway need to be resurfaced, what matters is cost-efficiency to fulfill the given service requirements.
(2) *Measurement*. Some of these objectives, like fairness, access, participation, or due process, are difficult to specify, hard to operationalize, and rarely directly observed in data.
(3) *Specification*. Moreover, some of the objectives of public services – such as teacher effectiveness and public safety – are initially under-specified, implicit, or vague. To what extent a government program meets these objectives depends, in part, on how these objectives are operationalized and measured in practice.
(4) *Trade-offs*. Finally, how to make tradeoffs – that is, how to determine the relative importance of such values, when they come into conflict with one another – likewise is under-specified and context-dependent.

Often, the best that AI developers – whether internal or external to the government – can do, by contrast, is to (1) "flatten values" by reducing dimensionality (Bueno de Mesquita 2019), (2) optimize for a measurable – but often flawed – proxy, (3) formalize and specify initially vague objectives without explicit public deliberation, (4) surface or hide trade-offs and conflicts between values that democratic systems have so far been unable to resolve.

These problems are familiar to both AI researchers and public management scholars. In public management – especially in New Public Management approaches to public administration – performance measurement is the key

[1] That is, design decisions are not made purely on empirical data and have consequences that can be assessed morally.
[2] Moore (1995). Public values are social, professional, ethical, and other values that are related directly to a public manager's role in serving the interest of the public (Moore 1995).

PROOFS

management tool, where a target (such as specific test scores) is taken as a proxy for a harder-to-measure objective (such as a rounded education) (Hood 1995). This incentivizes gaming through metric management (Eckhouse 2022). When a metric is turned into a measure, teachers teach the test, healthcare workers "upcode" their contribution (Li 2022), and much effort is expended on the performance of the metrics. In short, intended objectives are attained only speciously by gaming the metric. This is the sociological version of what AI researchers know as *reward hacking* on the alignment problem (Amodei et al. 2016): the model might optimize for an imperfectly correlated, misspecified, but easier to attain, objective, instead of the intended one.

These challenges – objectives that are multidimensional, complex, hard to measure, and – are not exclusive to the public sector. But in public sector work, these challenges loom large. Whereas the private sector tends to operate on well-defined metrics – such as revenue, number of items sold, or default rates of granted loans – articulating and specifying public service values is challenging already for seemingly simple services, such as issuing a driving license. Of course, licenses (especially if they serve an official identification function) should be issued only to individuals who produce valid documentation and pay a fee. No license should be issued based on forged documentation and, therefore, all presented documents must be carefully and rigorously audited. However, driver's licenses should be issued in a timely and convenient manner. These requirements create a tension between timeliness and fraud minimization. Detecting fraud is complicated and takes time because applicants' names change, occur in multiple spelling variants, and might include non-Western characters that are not recognized in the limited character encoding of legacy IT systems (McKenzie 2010). Consequently, digital identity systems, such as India's Aadhaar, are prone to substantial administrative errors, while still being compulsory for accessing many public services (Khera and Moharil 2024). To make matters worse, legislatures and administrations at different levels of government impose increasingly burdensome, fine-grained, and sometimes conflicting requirements on programs, degrading service quality and expediency. Given this complexity, it is generally difficult to specify all relevant values and objectives with the clarity required for engineering an AI system.

Such problems might be regarded as implementation challenges, but the issue of defining the system's objectives goes deeper. The five key design decisions of AI systems are profoundly normative, not technical (Coyle and Weller 2020). Key design decisions – such as what data to use, what inductive biases to assume, what objective to maximize, and how to present recommendations – could determine whether an AI will be a "force for social justice" or, more likely, entrench existing patterns of injustice and discrimination (Ludwig and Mullainathan 2021).

One deep dilemma raised in the design of AI systems involves mutually incompatible definitions of "fairness" (Kleinberg, Mullainathan, and

PROOFS

Raghavan 2016; Chouldechova 2017). In healthcare, a model that allocates scarce resources might be "fair" if it predicts equally well the potential to benefit a patient, regardless of age, gender, or race. Alternatively, the model might be "fair" if its prediction mistakes are equally likely regardless of age, gender, or race. But in almost all situations, a fairness objective can be met either in the first or second senses – but not in both. Which one should be prioritized? Public deliberation and legislative discourse rarely surface, let alone answer, such questions. Moreover, such system design questions are hard in a society where voters and politicians hold widely different beliefs and value systems.

Political systems have faced such hard, value-based questions before. Sometimes, these value conflicts are left to bureaucratic discretion. Other times, such questions have been answered through "incompletely theorized agreements," that is, by leaving some contentious issues unresolved or intentionally vague (Sunstein 1995). Artificial intelligence narrows the room for such incompletely theorized agreements. Artificial intelligence requires a new level of fine-grained specificity and pushes to the surface conflicts that did not previously arise in practice. Take the rules of the road: What lateral lane distance should heavy vehicles keep from cyclists? If a collision is inevitable, should a car collide with another car in front of it, or the car driving behind it? These questions were never issues that could be legislated, let alone encoded or enforced. But with autonomous vehicles, these questions arise and must be answered – by law or by code – every second of the day, and at scale (Goodall 2019). Whereas human drivers decide individually and intuitively how they drive, engineers make design decisions that ripple out across millions of vehicles in a fleet. They determine not only how one car drives, but how all cars drive. Autonomous vehicles, therefore, make even mundane driving behavior a potential policy issue. Otherwise, algorithms become policy.

Artificial intelligence now shifts and conceals discretion (Zouridis, van Eck, and Bovens 2020). For decades, information and communication technology (ICT) shifted discretion from a decentralized "street-level" to a consolidated "screen-level" of decision guidelines and regulations (Bovens and Zouridis 2002). Artificial intelligence further shifts discretion toward the *system*-level – and conceals it. A relatively small number of system designers now shape public service at scale and with limited oversight (Young, Bullock, and Lecy 2019). This makes the design of AI systems a point of powerful influence (see Section 4.3.2). Moreover, while the reduction of discretion eliminates *one* kind of accountability problem – administrative discretion threatens the rule of law as unelected career civil servants make arbitrary decisions – it creates new ones (see Section 4.3.3).

None of this is to argue that the public sector should not be using AI systems. The technology offers the prospect of better quality, more timely, and efficient services. Artificial intelligence may even ultimately reduce the cost of service delivery. There is good evidence of the gains provided by existing digital government systems, from Estonia's X-Road platform to India's Aadhaar.

PROOFS

Given that AI is a prediction technology, improved predictions offer opportunities to intervene in ways that improve outcomes for citizens.

Notwithstanding the various benefits of AI, the challenges identified previously hinder the responsible deployment of AI:

- First, public values that are unmeasured and under-specified risk being diminished. Public service values are multidimensional and latent. If AI is optimized primarily for outcomes that are easily measurable, those harder-to-measure valuable objectives – such as due process or fairness – risk falling by the wayside (Himmelreich, forthcoming).
- Second, AI requires hard, normative, value-based design decisions, including algorithmic fairness. If the political process fails to produce agreement on these issues, then decisions will be made outside of the political process, by private actors or through administrative discretion. This is a procedural problem and threatens democratic legitimacy (Meier 1997; Nabatchi, Goerdel, and Peffer 2011).
- Third, a valorization of AI as an "intelligent" technology might lead to AI exuberance: AI might be deployed aggressively in the public sector, often heralded as a cost-saving measure, even if the previous two problems are unsolved (Young et al. 2021).

Solving these problems requires institutional mechanisms and processes for accountability – the topics of Sections 4.3.2 and 4.3.3.

4.3.2 Challenge 2: Information and Power Asymmetry When Governments Outsource AI Innovation

To close the widening gap between technology and public service, governments are actively trying to build AI capacity. Instead of developing it internally, though, they often acquire it through contracts with third parties. Put differently, to understand how governments use AI, one must first understand how they buy it. In this subsection, we present a historically informed account of the information and power asymmetry between governments and private vendors from the perspective of the US federal government; nevertheless, other governments face similar challenges (OECD 2025).

Consider the case of the US, home to many of the world's leading AI companies. As of the date of the most updated federal report, January 23, 2025, the US government had documented 2,133 active AI use cases across 41 federal agencies. Among 1,147 cases that reported how they were developed, 40 percent were developed in-house, 42 percent under contract, and 18 percent through partnerships between agencies and contractors (see Figure 4.1). In other words, in-house development is less common than contractor involvement. A case in point is Elsa, a generative AI tool used by scientific reviewers and investigators at the US Food and Drug Administration (FDA), a federal regulatory agency with international influence. The customized AI tool, which operates in a high-security government cloud environment, launched

PROOFS

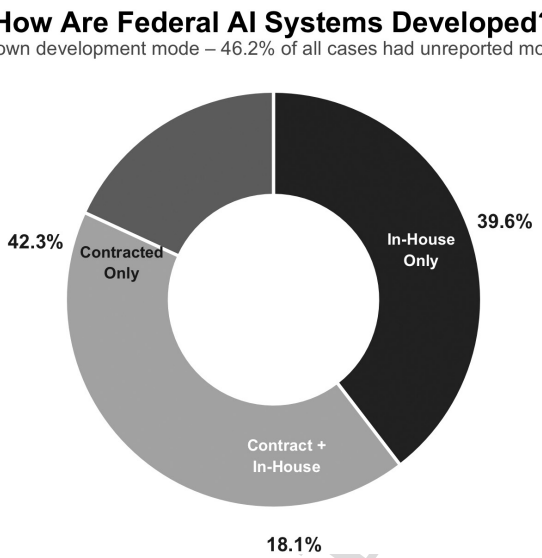


FIGURE 4.1 Share of federal AI use cases by development mode, among cases with known status (Mathur and Zimmerman 2024).

in June 2025. It was developed by Anthropic, the creator of the Claude AI platform, in partnership with Amazon Web Services, the leading cloud provider, and Deloitte, one of the largest legacy vendors in the government technology space (US Food and Drug Administration 2025). Moreover, there is a growing trend of federal government employees using public-sector versions of enterprise AI tools that are not specifically designed for their agencies. This expanding gray area matters because adoption costs are heavily discounted, at least temporarily, and AI use cases become more widely diffused. For example, following OpenAI and Anthropic, Google reached an agreement with the US General Services Administration in August 2025 to offer "Gemini for Government" at a total cost of about forty-seven cents for each agency for a year (US General Services Administration 2025).

Although not well documented, reliance on tech vendors for AI acquisition at the state and local level may be even more serious, given the deeper shortages of in-house expertise. For example, in November 2025, the state of Maryland announced plans to transform government services in areas such as child poverty, housing, and service modernization through two vendors: Anthropic, the AI developer noted earlier, and Percepta, which supplies engineering and technical staff (Maryland Office of the Governor 2025).

This reliance on contractors to modernize government is not unique to AI. The governance challenge long predates the arrival of AI.[3] Although the

[3] For general insights on the contracting state, Jae Yeon Kim thanks Alisha Holland and Gautam Nair (forthcoming) for sharing the introduction to their forthcoming edited volume.

PROOF

government designs policy, its technical implementation has increasingly been carried out by private contractors in areas ranging from waste management and health service delivery to community development and call center operations. The modern government, in this sense, is a contracting state. Moreover, the scale of the contracting state is staggering. While the number of federal civil service employees has remained relatively stable since the 1980s, at around two million,[4] the number of contract workers has increased to more than thrice that level (Light 1999, 2017). In 2023, according to Brookings' 2025 report, the number of federal civil service employees was about 2.1 million (Kamarck 2025). Meanwhile, contract employees numbered 5.22 million. For every one federal civil service employee, there were about two to three contract workers.

The New Deal era (1933–39) established many new federal agencies, often called the "alphabet agencies," such as the Federal Communications Commission and the Social Security Administration. This expanded the federal government's role in economic management, social welfare, and infrastructure development, in large part by regulating these domains through rulemaking. As the range of tasks the federal government needed to manage grew increasingly vast and complex, it could no longer rely solely on Congress's capacity to make rules. The Administrative Procedure Act of 1946 codified the procedures agencies must follow when exercising delegated rule-making authority.

In the decades that followed, these administrative powers were further delegated to private vendors, especially during the Reagan administration (1981–89) (Kettl 1993). As part of the broader privatization agenda, the administration promoted contracting as a way to reduce the size of government without limiting its scope. The same government that once sent people to the moon now struggles to build a functioning website, let alone develop its own AI capacity.

The hollowing out of the state's technical expertise is not accidental; it is a deliberate outcome. The Office of Management and Budget's Circular A-76, first issued in 1966 and revised multiple times, including during the Reagan administration, stated that "the Government should not compete with its citizens" (Office of Management and Budget 1999). In essence, if the federal government could buy it, it should not build it. As a result, many government programs became *publicly funded* but *privately operated*. More importantly, the state's emphasis on expertise shifted from developing technical skills and building systems in-house to hiring and managing contractors to perform the actual work. As government employees spend much of their time writing contracts and managing vendors, technical staff find it difficult to maintain, let

[4] Perhaps the exception is the second administration of US President Donald Trump. According to the Partnership for Public Service's Federal Harms Tracker, as of October 23, 2025, more than 211,000 civil servants had left the federal workforce due to firings, voluntary buyouts, or resignations. https://federalharmstracker.org.

PROOFS

alone improve, their skills. Those who wish to remain technical experts face incentives to leave government and pursue private-sector employment.

The operation of safety net programs illustrates the extent to which the US government depends on contractors to deliver high-impact services. Most programs are fully or primarily funded by the federal government but administered by state and local governments. In practice, however, the technical systems that support these programs are largely operated by private vendors. Eligibility systems and their digital interfaces, such as websites and mobile applications, are typically developed and maintained by contractors. Medicaid is a government-run health insurance program that serves roughly seventy-five million low-income Americans. As of April 1, 2023, one vendor, Deloitte, had built eligibility systems in twenty-five states. In other words, the applications of the fifty-three million Medicaid enrollees in these states – about 70 percent of all participants – are processed through Deloitte-built data systems. (Liss and Pradhan 2024). Street-level bureaucrats know their work, but their perspectives are bounded by their authority and jurisdiction (Simon 1947; Lipsky 1980). Caseworkers understand their individual cases, and program directors understand their specific programs, but each sees only part of the overall system. Ironically, it is often contractors, not state governments, who possess the most comprehensive knowledge of how these data systems function. This information asymmetry also creates a power asymmetry. Contracting relationships depend on information about the product, the vendor, and the task environment. The challenge is that searching for vendors, acquiring information, and negotiating contracts are all costly processes (Coase 1937; Williamson 1981).

Technically, public agencies award contracts to private firms, making the agencies the principals and the contractors their agents. In practice, however, states must rely on private vendors to understand how their safety net programs operate, including key measures of performance such as take-up rates, equity such as disparities in access, and integrity such as error rates and the prevalence of fraudulent cases. These indicators are essential for states to meet federal reporting and compliance requirements, including recent updates to work requirements under H.R. 1, known as the One Big Beautiful Bill Act, enacted during the second Trump administration. Moreover, when governments lack technical literacy or access to underlying data, they cannot evaluate vendor claims, benchmark performances, or audit the quality of reports produced by their contractors.

This imbalance creates conditions for vendor lock-in, where state and local governments become dependent on their vendors. They also find that the systems contractors build are costly and difficult to replace. Over time, these systems shift from tools of service delivery to instruments of control by agents over their principals. Based on historical data on federal mainframe computer procurement, Greenstein (1993) shows that the incumbent contractor, in this case IBM, enjoyed an advantage even after controlling for other factors such as

PROOFS

vendor–buyer matching. Similarly, Brunjes (2020), drawing on nearly 25,000 federal contracts, argues that competition does not deliver the expected benefits of cost savings, innovation, or improved accountability.

Artificial intelligence is new, but the governing challenge it poses is not. Managing vendors has long been a central problem of the delegated state (Morgan and Campbell 2011). Asymmetric relationships, power imbalances, and information asymmetries have long structured interactions between governments and private actors. What is new is the scale and speed of technological development and the extent to which it intensifies these asymmetries.

Procuring AI faces many of the same challenges that procuring technology has encountered over the past few decades. Kettl (1993, p. 204) warned that excessive dependence on contractors weakens state capacity and erodes governmental accountability. He noted that sharing expertise between government and private partners is inevitable, since the government cannot build everything and the domains it governs continue to expand. Yet he cautioned that sharing power is a different matter: "Government must have an independent, smart-buyer capacity if it is to govern."

Decades later, in the era of AI, Kettl's warning remains just as relevant. To be fair to the proponents of contracting, they were aware of this policy tool's potential dangers and limitations. The Reagan administration's 1988 report *Privatization: Toward More Effective Government*, chaired by David Linowes, a political economist who advised four presidents, praised the virtues of the free market and privatization. To their credit, however, the authors also acknowledged the risk that contractors could become entrenched interest groups (Anderson and Linowes 1988, p. 244). To prevent such lock-in from occurring, the President's Commission on Privatization (1987–88) identified several conditions under which contracting can be effective. Meeting these conditions, however, is generally difficult and especially challenging in the case of emerging technologies like AI. Effective contracting assumes that delivery terms are clear, yet defining precise deliverables is rarely possible when the underlying technology is changing, and its use cases vary widely. It also depends on the presence of competition, which is often limited in technology markets (Wu 2018). This concern is amplified in the AI industry, where control over foundational models, advanced semiconductors, and cloud infrastructure is concentrated among a small number of firms. However, it is worth noting that there is a large number of new AI start-ups, especially in the application layer. Nonetheless, what these start-ups can do in that space is constrained by AI infrastructure companies. The assumption that contractors do not need to undertake large upfront capital commitment is similarly unrealistic, as technology development typically involves substantial upfront investment, often financed through venture capital (Mallaby 2022). This dynamic is even more pronounced in the current AI landscape, where so-called LLMs

PROOFS

Tucker_9781047774734_C004.indd 120 15-09-2026 08:55:42

demand massive capital outlays to build data centers and secure high-performance computing infrastructure. Start-ups can build application-layer products on top of these models, but their strategic choices are constrained. Finally, contracting works best when agreements are negotiated, reviewed, and renewed regularly. In practice, however, at least two problems arise. First, it is difficult to monitor whether vendors deliver what contracts specify. This challenge is especially acute for AI systems, given governments' limited technical expertise and the proprietary nature of many commercial AI technologies. Under these conditions, vendors tend to prioritize cost reduction over outcome quality (Hart, Shleifer, and Vishny 1997). Second, incumbent contractors, often called legacy vendors, enjoy strong advantages through institutional knowledge, established relationships, and past performance records, making it difficult for new entrants to compete.

The US case is not an outlier. The contracting state is a broad, global phenomenon. According to 2023 OECD statistics, governments spent on average about 30 percent of their total expenditures on procurement (OECD 2025). The US share was actually lower than the average, at 25.4 percent, yet the scale of federal contracting remains immense given the size of the US government and its contracts.

4.3.3 Challenge 3: The Crisis of Accountability

Accountability is a procedural cornerstone of democratic governance. Democratic accountability includes both the lived practice and the formal rules by which some agents evaluate the conduct of others. For example, citizens evaluate how well their elected representatives have performed and, therefore, whether they should be reelected (Bovens 2007). Citizens hold representatives accountable *informally*, for example, by expressing discontent on social media, but also *formally*, by voting for another candidate next time. Accountability is a relation between agents governed by such informal normative standards, which we call "responsibility practices," and formal institutions, which we call "accountability frameworks," that codify and enable the enforcement of these standards (Fleisher et al. 2025). Both the responsibility practices and the accountability frameworks are evolving. Often, the frameworks are slow to catch up with a rapidly changing practice. Moreover, accountability involves not only the relationship between agents but also the virtue of being open to scrutiny and to explain decisions (Lechterman 2024).

Artificial intelligence changes the dynamics of accountability (Cooper et al. 2022). Schematically, accountability – both the informal practices and the formal frameworks of accountability – requires transparency. To demand answers, we have to know to whom questions should be directed and for what conduct. But the development of AI is not transparent, and AI itself is often not fully explainable.

PROOFS

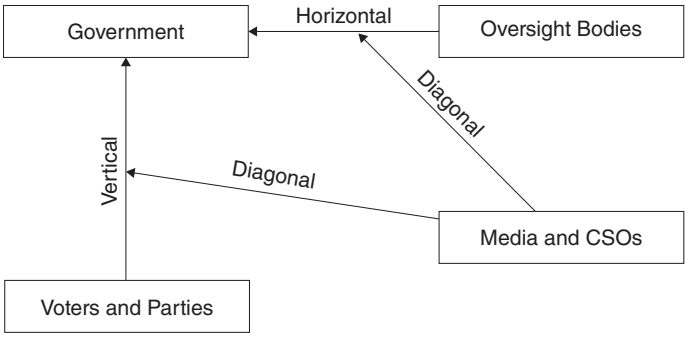


FIGURE 4.2 Traditional accountability relationships in public administration (Lührmann, Marquardt, and Mechkova 2020).

First, the development of AI is a complex, large-scale, multi-actor collaboration. Having many hands involved in the pipeline of AI development, procurement, and deployment makes it practically difficult to determine the set of actors who owe accountability. Moreover, AI development involves information and power asymmetries that further obscure who ought to be held accountable for what. This potentially impedes accountability both as an informal practice and as an accountability framework. For example, legal liability is harder to establish as the situation gets more complex. Moreover, AI models themselves are opaque. The output and the decisions of foundation models are not easily explainable. These two points taken together allow human actors to avoid accountability by using AI as a scapegoat. The nontransparency and opacity of AI can allow humans to misrepresent their own role in a decision ("Don't blame me, the AI did it"). Finally, even if the set of accountable actors *can* be identified, AI requires novel normative standards according to which accountability-seeking actors hold other actors accountable. For example, the decision of when to use AI – given the opacity of AI models – is a decision made under severe uncertainty. What are acceptable and unacceptable ways for accountable actors to make decisions with such high uncertainty?

In this section, we explain how the introduction of AI changes the dynamics of accountability in government and public administration (see Figure 4.2). The fragmented responsibility chain ("many hands" problem) and the tendency to scapegoat computers and software hinder vertical accountability, or the ability of legislative oversight or electoral responsiveness. The "black-box" nature of closed-source AI and the asymmetry in technical expertise make auditing highly difficult. It weakens horizontal accountability by independent regulators and weakens diagonal accountability by limiting the ability of civil society and journalists to push back against the harmful deployment of the technology. Furthermore, most AI models are proprietary and protected by trade secrets. In theory, governments can require vendors to increase transparency and accountability. They can ask vendors to disclose how models

PROOFS

were trained and to permit ongoing monitoring and auditing of performance (Coglianese and Crum 2025). In practice, however, this is difficult when governments do not even know what questions to ask or lack the technical staff needed to carry out these functions. We supplement theoretical discussions with real-world case studies of how existing accountability mechanisms failed to prevent harms caused by public sector use of AI.

4.3.3.1 Vertical Accountability

The complexities of modern governments require elected officials to delegate responsibilities to bureaucrats who have specialized skills or expertise to serve the public. Vertical accountability within government depends on principals monitoring agents to whom they delegate tasks, ideally with each level monitoring the level below it. Chains of responsibility typically involve the public delegating to elected officials, who in turn delegate to agency heads, to senior managers, and to street-level bureaucrats. Accountability flows upward through this chain via monitoring and the threat of sanction (e.g., bureaucrats getting terminated, politicians getting voted out of office). Private vendors are not depicted in the diagram as subjects of accountability, since they are hired and not elected or appointed.

When governments deploy AI in the public sector, they further delegate responsibility to computer software systems. Although the popular imagination would like to think of an AI system as a unitary actor, in reality, it is the product of "many hands." As a result, deploying AI significantly complicates the chain of accountability for vertical accountability – particularly when the government procures AI from a third party. Suzanne Mettler (2011) showed that, when citizens are unaware they are receiving government support because the state is hidden or, in her words, submerged, the government fails to receive the credit it deserves. The contracting state represents the opposite problem. When citizens do not realize that the difficulties they face in interacting with the state stem from contractors, it becomes difficult for them to know whom to hold accountable. Although James et al. (2016) show that information about contracting out does not reduce citizens' attribution of blame to politicians, it is unrealistic to assume that most people know how much government services depend on vendors, especially in the technology sector. As described in Section 4.3.1, computer programmers design algorithms to optimize for goals that may be far removed from what the public (the ultimate principal) wants.

Recent lawsuits in the United States also suggest that government agencies using algorithmic systems failed to follow the established administrative due process. For example, in the case of the Michigan Integrated Data Automated System (MiDAS), deployed to address unemployment fraud, the government wrongfully accused 34,000 individuals of committing it (Charette 2018). Although such errors could have happened previously as well, bureaucrats were supposed to send notices to appeal to individuals' addresses. But with the automated system, the notices did not arrive, which took away the right to

PROOFS

administrative appeal and assessment. The case illustrated that private companies involved in developing automated systems deployed by public authorities could potentially be held legally liable for harm resulting from those systems, even if they did not directly oversee the systems' use (Cortez and Maslej 2023).

This MiDAS case is an example of how formal accountability frameworks, like the law, can respond to new challenges that AI presents. Moreover, it is an example of how formal accountability frameworks can operate alongside informal responsibility practices, as journalistic reports on MiDAS accompanied the litigation. But formal accountability frameworks do not always align with informal responsibility practices. On other AI-based tools, courts have ruled that private companies can, under the trade secrets doctrine, refuse to explain their algorithms. In such cases, courts have ruled that due process clauses require the government agency to provide the information the government relied on to make its decisions (Coglianese 2023).

Despite the fact that AI systems are a product of many hands, human actors along the accountability chain often scapegoat AI. It may be easier to assign blame to a single villain than to multiple actors or systems. For instance, when students in the UK protested against the A-level grading algorithm that disadvantaged pupils from under-resourced, low-performing schools in undergraduate admissions, their rallying cry was "fuck the algorithm" (Kolkman 2020). While the algorithm used was faulty, racial and class inequality within the education system contributed to the training data that fed into it. Likewise, the state government of Indiana blamed software for its welfare automation scandal, in which an IBM-designed system led to one million benefit denials over a three-year period. But when the program was first implemented, the state lauded the contract with IBM as an effective means to reduce welfare fraud and bureaucratic bloat. In an effort to avoid blame for privatizing its welfare services, which led to a surge in benefit denials, the state of Indiana successfully sued IBM for $78 million (Callahan and Davies 2017). In both cases, it was much easier to blame the software for harming the public rather than to shift the focus to the human institutions that make decisions about the design and deployment of algorithms.

The examples discussed earlier involve narrow-purpose algorithms where it is feasible to pinpoint the causes of faulty predictions. Auditing such algorithms and their underlying training data – which are often more transparent and explainable than human decision-making processes – allows developers to identify variables that are poor proxies for the social phenomenon they are meant to represent or for flawed modeling assumptions (Obermeyer et al. 2019). However, the general-purpose foundation models that governments are increasingly relying on are vastly more complex and less explainable when they make predictions. Commercial vendors typically provide only API access while withholding model weights, training data, and internal documentation. According to Stanford University's Center for Research on Foundation

PROOFS

Models' May 2024 Foundation Model Transparency Index, the median open developer they studied had models that had higher transparency scores than the median closed developer on all but five of the twenty-three subdomains analyzed (Bommasani et al. 2024). Not having access to internal documentation about how the model was developed (e.g., code, training data, and the training process) limits the precision of audits and makes the auditing process inefficient (Casper et al. 2024).

4.3.3.2 Horizontal Accountability

Governments' eroded technical capacity, as discussed in Section 4.3.2, and the proprietary opacity of closed-source foundation models challenge horizontal accountability. Horizontal accountability institutions, such as auditing bodies, legislative committees, and administrative courts, were designed to oversee government agencies through independent evaluations and legal authority (O'Donnell 1998; Bovens 2007). But in many countries, governments procure closed-source AI systems without having the internal expertise to evaluate vendor claims or without setting up oversight mechanisms to overhaul AI systems when they cause harm.

This might be a case in which the formal accountability frameworks are in the process of catching up with informal responsibility practices. Auditing bodies, legislative committees, and administrative courts are updating the formal accountability frameworks available to them to exercise horizontal accountability despite the new challenges introduced by AI. For example, the US GAO, in a 2023 report, acknowledged that third-party assessments and audits are important to ensure AI systems adopted for governmental use are safe. But, it said, "a critical mass of workforce expertise is needed to enable federal agencies to accelerate the delivery and adoption of AI" (US Government Accountability Office 2023). Similarly, the UK National Audit Office's 2024 survey found that 70 percent of government bodies identified difficulties in finding and retaining staff with AI skills as a critical barrier to AI adoption (UK National Audit Office 2024). This capacity crisis means that, even when oversight bodies theoretically possess audit authority, they lack the technical sophistication to exercise it effectively. Evaluating AI systems requires technical competence to assess whether integrating a specific AI application is appropriate for an existing workflow. Such competence is also required to vet the quality of the training data and to interpret model performance metrics, which require both technical and institutional expertise (Green 2022, p. 20; Aldemir and Uçma Uysal 2024).

4.3.3.3 Diagonal Accountability

Diagonal accountability, through scrutiny and advocacy by investigative journalists, civil society organizations, and academic researchers (Lührmann et al. 2020), faces many of the same problems as governmental oversight bodies. Ideally, these nongovernmental actors enhance democratic governance by

PROOFS

exposing government failures, amplifying citizen grievances, and pressuring officials to redress or mitigate AI harms. Indeed, many of the harmful AI incidents we discuss in this chapter were first brought to light by journalists, digital rights groups, and academic researchers. But technical opacity and legal threats have increasingly made these actors' work more difficult in recent years.

The technical complexity of AI systems creates substantial barriers for the media and civil society organizations. Unlike traditional bureaucratic decisions that leave paper trails and can be reconstructed through interviews and document review, algorithmic decision-making often operates as a "black box" that resists external interpretation (Busuioc 2021). These challenges are amplified by newer generative AI systems, such as the LLMs with which governments are experimenting. Earlier algorithmic systems, such as Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) or MiDAS, were designed for specific, bounded tasks: They use a limited set of variables to predict a limited set of outcomes. In contrast, foundation models are trained on vast corpora of data and can be adapted to an essentially unlimited range of downstream tasks (Bommasani et al. 2021). This is an area where the formal accountability mechanisms have not yet responded to the new challenges of nontransparency and opacity that AI introduced.

Furthermore, academic researchers and civil society groups auditing AI and algorithms have experienced increasing legal threats from tech companies. In one prominent case in October 2020, Facebook sent a cease and desist letter to two New York University researchers, demanding they shut down a browser extension that allowed volunteers to share data about political ads they were shown on the platform. When the researchers refused, in August 2021, the platform terminated the Facebook accounts of the researchers (Whitford 2021). Similarly, in 2021, the non-profit organization AlgorithmWatch was forced to halt its work studying Instagram's algorithms after the company threatened to take legal action against it (Cameron 2021). As the boundary between tech companies and the government blurs in the US context, the government itself has threatened researchers and civil society groups. Directed by the Department of Government Efficiency (DOGE), the US National Science Foundation, an independent federal agency, canceled numerous grants related to studying misinformation and disinformation on social media platforms (Myers 2025). In a more extreme example, in December 2025, the US State Department imposed visa bans on Imran Ahmed, the CEO of the Center for Countering Digital Hate, and four other Europeans, accusing them of coercing tech platforms to censor American viewpoints (Kavi 2025). Given the influence of tech companies in providing AI services to governments, these actions may produce a chilling effect that discourages researchers and civil society groups from auditing governmental AI systems developed by private entities.

In short, AI disrupts all three relations of democratic accountability: vertical, horizontal, and diagonal. Artificial intelligence makes responsibility chains

PROOFS

opaque and diffuse, which requires transparency and identifiable agents. The many-hands problem lets human decision-makers scapegoat software; trade secret protections let vendors refuse to explain their systems; and the technical capacity deficit leaves oversight bodies without the expertise to audit what they are nominally empowered to scrutinize. Informal responsibility practices, as well as the accountability frameworks along the different relations of responsibility, need to adapt to a changing environment. Responsibility practices adapt faster with citizens voicing discontent when officials try to scapegoat an algorithm. Also, formal accountability frameworks – through courts, audit institutions, and legislative committees – are beginning to adapt, as the MiDAS ruling and GAO reports illustrate. But they are catching up slowly, and the shift from narrow-purpose algorithms to general-purpose foundation models is widening the gap between what accountability practices require and what current formal accountability frameworks can deliver. The upshot is that the deficit of accountability around public-sector AI is primarily institutional, not technical. The institutional deficits can be overcome by establishing transparency requirements, developing technical capacity, and clarifying the goals and values in governance frameworks before delegating consequential decisions to systems that, without such changes, undermine responsibility.

4.4 Research and Policy Agenda: Improving Democratic Accountability of Governmental AI

The challenges we outlined in Section 4.3 are formidable, but we argue that the proposed solutions are feasible. Addressing them requires coordinated efforts from researchers, policymakers, and practitioners to develop both a deeper empirical understanding of the use of AI by government and practical institutional reforms. This section outlines a research and policy agenda organized around four interconnected priorities.

First, we argue that transparency serves as the foundational prerequisite for all other reforms. Without systematic documentation of the AI systems governments deploy, for what purposes, and with what safeguards, neither researchers nor citizens can meaningfully evaluate whether the use of AI serves the public interest. We discuss both formal transparency mechanisms, such as algorithm registries, and alternative approaches that researchers must employ when proactive disclosure fails. Second, we examine how governments can improve the design and deployment of AI systems to better align with public service values, including frameworks for defining objectives, managing error rates, and ensuring meaningful avenues for redress when algorithmic decisions cause harm. Third, we argue that to rebuild state capacity in the AI space would require attracting technical talent to the public sector and creating work environments that empower experts to focus on mission delivery rather than procedural compliance. Furthermore, governments should leverage procurement as a tool for accountability rather than a source of vendor lock-in.

PROOFS

Finally, we assert that public policy research should focus on studying what institutional mechanisms – not merely what responsible AI systems – would ensure safe deployment of the technology. Throughout, we emphasize that political scientists and public administration scholars have essential roles to play in advancing both empirical research and institutional innovation.

4.4.1 Improving Transparency

The most basic prerequisite for advancing research on the use of governmental AI is greater transparency. Before researchers can evaluate whether algorithmic systems serve the public interest, they must first know which tools are being deployed, by which agencies, for what purposes, from which vendors, and with what oversight. Yet many governments fail to disclose this basic level of information. Greater transparency would require establishing and maintaining publicly accessible registries documenting government AI systems in detail. Short of such measures, researchers must rely on "unorthodox" approaches, such as using freedom of information laws to access information and collaborating with investigative journalists, digital rights groups, and citizen scientists. Transparency does not automatically lead to accountability when addressing harms caused by AI systems, but it is a necessary ingredient.

The current state of government AI transparency is lacking, although this is recognized by policymakers. For example, the EU's (voluntary) AI Code of Practice, introduced in mid-2025, sets out best practices in terms of transparency for developers of generative AI models (European Commission 2025a). Algorithm registries, publicly accessible databases cataloging government AI systems, have also emerged as a leading transparency mechanism, pioneered by cities such as Amsterdam and Helsinki and subsequently adopted by national governments, including those of the Netherlands and the United Kingdom. But empirical research suggests that existing registries often have basic information gaps. A 2024 report from the Global Partnership on Artificial Intelligence (GPAI) identified seventy-five active repositories of publicly available algorithms across five continents – a surprisingly small number given the proliferation of AI in government as discussed in Section 4.2. More troubling, many of these repositories catalog only a limited number of government AI systems, and most of these repositories provide only sparse information, making it impossible for researchers to evaluate the underlying AI systems (Global Partnership on Artificial Intelligence 2024). The GPAI highlighted that the US federal government hosts one of the more comprehensive repositories. Nevertheless, Lawrence, Cui, and Ho (2023) found that nearly half of 220 federal agencies failed to publicly publish AI use-case inventories, even when they had demonstrable machine-learning deployments. Even the US GAO has identified persistent inconsistencies in reporting standards across agencies and has recommended improvements (US Government Accountability Office 2023).

PROOF

What should scholars of public policy or political science do in light of this transparency gap? We argue that researchers must employ more creative approaches. Freedom of information laws, also known as public records laws, have historically served as essential tools for investigating government operations (Bloch-Wehba 2021). For instance, ProPublica's landmark investigation of the COMPAS algorithm using public records requests demonstrated what determined researchers can accomplish despite these obstacles (Larson et al. 2016). Nevertheless, because of trade secret protections claimed by private vendors, freedom of information laws are not a panacea (Wachter, Mittelstadt, and Floridi 2017). As discussed in Section 4.3.3, researchers and journalists cannot comprehensively audit closed-source AI systems and must rely on assumptions based on limited data or information about the models (Garrett 2020).

Another creative avenue for research is collaborations with investigative journalists, civil society groups, and citizen scientists. Investigative journalists have brought to light harms caused by public sector AI in several prominent cases and, in some cases, have shared data and other artifacts that support further analyses. For instance, the COMPAS data that ProPublica has made publicly available after the release of their report has become a canonical benchmark dataset in algorithmic fairness, with more than 81 scholarly articles utilizing it as of 2022 (Fabris et al. 2022).

Collaborations between academic researchers and digital rights organizations have significantly advanced transparency in public sector AI by combining research expertise with advocacy. For example, the Ada Lovelace Institute worked directly with UK governmental bodies to develop meaningful transparency mechanisms. They published foundational research on public registries, documenting algorithmic decision-making systems across local authorities that informed the UK's Algorithmic Transparency Recording Standard (Ada Lovelace Institute 2020).

Citizen science offers another pathway to increasing transparency, particularly when dealing with opaque commercial AI systems. Public donations of data allow researchers to collect otherwise proprietary data. Volunteers, through projects like AlgorithmWatch and the National Internet Observatory, installed browser plug-ins and donated their data concerning how social media algorithms curated their internet experience (Kayser-Bril 2024; National Internet Observatory, Northeastern University 2024). The public can also contribute to the evaluation of AI systems through crowdsourcing of harmful AI incidents (AI Incident Database 2024) or red-teaming challenges to evaluate public-facing LLMs (Humane Intelligence 2024). These types of participatory evaluation could be extended to governmental uses of AI.

Transparency, while insufficient on its own, remains a prerequisite for meaningful accountability in governmental AI. Without knowing which AI systems are deployed, by which agencies, for what purposes, and with what safeguards, researchers cannot evaluate whether these applications serve the public interest. Moving beyond this basic line of inquiry, researchers can then

PROOFS

begin to ask which transparency practices most effectively improve public services and governance (Kosack and Fung 2014).

4.4.2 Aligning AI with Public Service Values

Careful thought should be given to the definition and engineering of system objectives before deployment, to ensure they align with public service values. This includes explicit consideration of the key design decisions involved in developing AI, such as

- decision-makers' tolerance for false positives and false negatives (one dimension of fairness, as discussed above);
- the weighing of equity versus efficiency;
- whether to include data on minor offenses, such as fare dodging in training data; or
- whether to choose a simpler, more explainable model, over a more complex, black-box one.

Moreover, even a small error rate can affect very many people when large numbers are involved. There are well-documented differences in outcomes among population groups in systems such as facial recognition. Similarly, data bias is a well-known challenge that can be subtle and hard to redress. For example, one study found that, among claimants for disability payments in the US, women were far more likely to be rejected by the algorithm. The reason turned out to be differential responses to questions about physical mobility – the ability to do laundry or go shopping. Women do these tasks more often than men (Low and Pistaferri 2025).

At the design and pre-deployment stage, a practical framework for ensuring public service aims are achieved should include questions about data sources and outcome metrics. An ongoing research problem is deciding how to measure and specify public service values and how to trade off conflicting values against one another. The pre-deployment stage should also consider explicitly the threshold error rates for both false positives and false negatives, recognizing that there is a trade-off between them. And the approach should be more cautious in domains that are high impact (such as welfare payments or policing) and/or where the incentives of the state and the individual are not aligned (as accountable governments should be wary of overriding the aims and wishes of individual citizens, even if ultimately they conclude it is the right thing to do).

Satisfying these criteria would usefully involve the participation of both the groups affected by the algorithmic decisions and of the officials deploying the system. Artificial intelligence systems are trained on codified dataset, which, by definition, omit the contextual know-how and experience of humans. If human judgment is being removed in downstream decisions by the use of AI, it is all the more important to involve it as much as possible in the upstream

PROOFS

design. The prior experience of digital government is that effective systems tap into human expertise before deployment (Pahlka 2023). Artificial intelligence and human judgment might complement each other. But human judgment is fallible, and the difficulty of key design decisions makes algorithms fragile (Ludwig and Mullainathan 2021). How human and AI can be combined is thus an open problem of human–computer interaction, psychology, and behavioral science.

The legitimacy of AI use in public service ultimately depends on creating transparent, straightforward administrative processes for correcting errors; individuals should not need to go to the civil justice system for redress. Although the use of AI presents opportunities to improve and better tailor public services, the problems and challenges being addressed by governments are intricate: They cannot be solved algorithmically (Turing 1936). Society is complex; life and politics are messy. Algorithmic government needs to ensure that human and political judgment retain their place.

4.4.3 Rebuilding State Capacity

Addressing the power and information asymmetry between the state and private vendors requires renewed scholarly attention to state capacity, understood broadly as the state's ability to solve public problems. Building such capacity depends on the government's ability to recruit the right people and place them in organizational environments that cultivate expertise and support the state's role as a smart buyer. Yet the process through which these conditions are constructed and managed remains under-theorized and under-studied. This gap underscores that advancing research on state capacity requires moving beyond institutional arrangements to examine capacity as an organizational achievement, particularly as a function of personnel management practices (Kapur 2020, p. 50). More specifically, this perspective points to a research agenda that centers on personnel management and organizational design as key explanatory variables in studies of state capacity in AI governance.

At least two challenges in this regard merit closer and more systematic scholarly attention. The first issue is hiring. As Ho (2023, p. 3) emphasized in his testimony presented to the US Senate Committee on Homeland Security and Governmental Affairs, "government cannot govern AI if it does not understand AI." Therefore, "getting technical talent into the federal workforce is … the biggest obstacle to the US government's internal adoption of AI." According to the 2023 report from Stanford's Institute for Human-Centered Artificial Intelligence (HAI), 65 percent of AI PhDs work in industry, 28 percent in academia, and fewer than 2 percent in government (Maslej et al. 2023). This imbalance raises a set of research questions about how labor market dynamics shape the state's ability to attract and retain technical talent. It also invites inquiry into how these constraints affect the government's capacity

PROOFS

to act as a "smart buyer" (Kettl 1993). Equally important, the protracted and procedurally rigid nature of federal hiring processes warrants scholarly inquiry into how administrative timelines and recruitment practices function as bottlenecks in capacity building. These frictions are all different aspects of capacity-building challenges, which have implications for how quickly and effectively governments can respond to emerging technologies, such as AI.

The second issue concerns the work environment in which technical expertise is expected to operate. Procurement has long been a challenge, but, rather than addressing it by bringing in better talent, the American administrative state has often tried to fix it by adding more rules and regulations. What legal scholar Nicholas Bagley called "the procedure fetish" is also evident here. Jennifer Pahlka is the founder of Code for America and a former US Deputy Chief Technology Officer under the administration of President Barack Obama. She helped establish the United States Digital Service, which brought technical talent into the federal government. Pahlka (2023, p. 91) recounted that agencies have been made accountable to the *process* rather than the *outcome*. She further observed that governments have developed increasingly elaborate procedures for procuring digital services, sometimes rendering the procurement process more complex than the technical work for which it is being procured (2023, p. 109). These observations invite scholarly inquiry into how procedural compliance shapes the everyday work environments of public-sector technologists and procurement professionals. Even when governments succeed in recruiting technical talent, organizational settings that privilege rule-following over problem-solving may constrain the state's capacity to learn, adapt, and exercise judgment (Honig 2024).

Procurement is a key challenge, yet it remains one of the few levers the government has over contractors through its buying power. Once a contract is signed, it becomes difficult to modify the terms or switch vendors. Even so, the contracting process is a window of opportunity. During negotiations, the government can include provisions that promote transparency and accountability from AI vendors. As Coglianese and Crum (2025) argue, the better way to regulate AI is not through prescriptive "guardrails," which are fixed and rigid. Such an approach is ill-suited to managing AI risk, since the technology continues to evolve and its applications vary widely, even when based on the same foundational model. Instead, Coglianese and Crum propose a "leash," or management-based, approach that requires firms and, in the procurement context, contractors to provide continuous oversight of their AI products and services. In practice, this could take the form of smarter contracts that require vendors to disclose how their models are trained, what data and benchmarks they use, and where these models perform best or worst. In the EU, the AI Act, adopted in 2024, is complemented by model contractual clauses developed for public buyers (European Commission 2025b). In 2025, Japan also issued government guidance for

PROOFS

generative AI procurement that provides model contract considerations (Digital Agency, Government of Japan 2025). Such contracts function as transparency-enhancing mechanisms (Kosack and Fung 2014). They can allow agencies to test and learn from these systems using their own data before committing to a full purchase. Fundamentally, this is also tied to the state capacity problem, since it presumes that the government knows how to negotiate responsible AI provisions with its vendors.

Yet negotiation requires more than expertise; it also requires bargaining power. Governments are often contracting with powerful AI companies, whose dominance has not, to date, been tempered by legislation such as the EU's Digital Markets Act or by legal challenge by the US Department of Justice. In November 2023, the City of San José, located in the heart of Silicon Valley, and roughly fifty founding agencies established the GovAI Coalition. The city was motivated to launch this initiative after encountering difficulties in obtaining information from its AI vendors. The Coalition operates seven committees, one of which focuses on procurement and develops governance frameworks for AI integration. This procurement committee is open only to government and civil society organizations and explicitly excludes vendors. It facilitates information sharing, develops best practices, and enables member governments to strengthen their position through cooperative procurement. For example, while a small local government may lack the bargaining power of a large one, it can benefit by joining or "piggybacking" onto cooperative procurements initiated by the latter (City of San José 2025). Researchers could investigate what forms of coalition-building could best enhance relatively weaker government agencies' collective bargaining power.

Both expertise and bargaining power are essential for the government to act as a smart buyer. Cooperative purchasing can save administrative costs, time, and effort. Yet, because the initial contract becomes the boilerplate that other members will follow with only minor addenda, how that first agreement is negotiated is critical and depends heavily on expertise. Civil society actors, including academic researchers, can play an active role. They can help state and local governments embed privacy and responsible AI terms related to transparency, fairness, and accountability in their contracts. They can also assess these practices and share their findings publicly. Doing so provides agencies with additional capacity and incentives to follow recommended standards (Kim and Kesari 2026).

More generally, procurement is an understudied governance mechanism in political science through which private actors accumulate power and exert influence over public governance. Scholars should devote greater attention to this mechanism, especially given that it is the primary channel through which AI enters government and generates downstream consequences. An important avenue for future research is how to build "smart buyer" capacity (Kettl 1993) amid the gap between the rapid AI development and slower organizational adaptation in government.

PROOFS

4.4.4 Building Accountable Institutions

Artificial intelligence generates benefits but also poses risks to individuals, organizations, and society as a whole by neglecting core public values, such as transparency, fairness, and equity (Wirtz, Weyerer, and Geyer 2019; Madan and Ashok 2023), or by changing the dynamics of public values (Kerkhoff 2016), as in the case of accountability.

Organizational changes in administrative contexts have often redefined accountability in public administration (Erkkilä 2007). It is one of the reasons why public administration scholars consider accountability as an ever-expanding concept (Mulgan 2000). At its core, accountability refers to the elected officials' and public servants' obligation to inform, explain, and justify their actions (Bovens 2007). However, who is accountable to whom, for what, and why they are accountable have changed over the years because major reforms reshape existing patterns of influence and accountability within government organizational structures and cultures, leading to the emergence of new dynamics (Christensen and Lægreid 2016).

Today, the use of AI systems results in two major changes with implications for accountability. First, to make the design decisions necessary in developing and deploying AI, technologies and AI-based tools are centralizing discretion, and moving away from citizen-facing, street-level bureaucrats, such as police officers and public works personnel (Lipsky 1980). Traditionally, street-level bureaucrats could interact with citizens to get the necessary information, reflect on the situation, and apply the rules and standards to provide the service. The expertise of bureaucrats helped interpret the policies to account for the different situations that citizens experience. Street-level bureaucrats were held to account in different forums to ensure their actions were within the purview of law. However, as AI systems replace street-level bureaucrats, the systems are not only limiting their discretion but also shifting the balance of power from bureaucrats to AI developers – who are often in the private sector. The increasing level of automation, especially using machine learning techniques, has resulted in the consideration of AI as an "agent embedded in an organizational context that executes tasks using stochastic, or non-deterministic, approaches" (Bullock, Young, and Wang 2020). This has led to the term "artificial discretion" (Young, Bullock and Lecy 2019) Further, because the use of AI formalizes the decision-making process while reducing transparency, it creates an accountability gap that raises a crucial question: Who is responsible for the decisions?

Second, accountability challenges also emerge when the source of decision-making authority is no longer a single individual. The development of AI-based tools requires the involvement of multiple actors. For example, algorithms are developed by systems designers or technology experts, while the training data is shared by data analysts. These actors are responsible for ensuring the quality of inputs used in developing and designing the automated

PROOFS

systems (Benbya, Davenport, and Pachidi 2020; Zouridis, van Eck, and Bovens 2020). If the AI systems are developed by government agencies, holding an actor accountable can be challenging. When private organizations develop the algorithmic system for government agencies, they are also expected to adhere to certain standards and policies in the development of algorithmic systems. But the self-learning properties of AI combined with the proprietary information may make it difficult to understand the decision outputs in some cases (Novelli, Taddeo, and Floridi 2023). As a result, public servants who use an AI system may find it challenging to explain and justify decisions (Schiff, Schiff, and Pierson 2021).

Third, accountability involves a system of review and redress. The AI systems will make mistakes intentionally or unintentionally. For example, the biggest corporate scandal in recent UK history was the Horizon case, involving the Post Office. Horizon was a computer system installed by the computer services company Fujitsu in every post office in the country. The system had coding errors that meant that every night when it added up the day's takings, it showed shortfalls, implying that postmasters were committing fraud. The system's designers and company executives knew about the errors from early on. They nevertheless pressed ahead with falsely prosecuting individuals. The courts believed the computer, and the individuals running a post office were up against the might of two large corporations. Some went to prison, some died by suicide. Many had their careers and reputations destroyed. Most convictions have since been quashed, and a redress scheme is still active. (Wallis 2022). This was a particularly egregious case, but few people would now accept the assumption that the computer is always right. Artificial intelligence systems are, notoriously, prone to hallucination.

Although there are discussions on accountability in relation to the challenges brought about by AI, as shown earlier, there is limited empirical knowledge on the specific changes in accountability dynamics, as well as on how government organizations are addressing the problem. In this respect, studies on AI accountability have largely focused on responsible design in the development and implementation of AI technologies. However, in public administration, the use of AI adds new challenges to accountability that, as previous public administration literature has made evident, require good governance practices that promote the creation of public value.

In addition, the AI literature rarely identifies the specific changes in accountability dynamics across different types of accountability and from the perspectives of different stakeholders, such as street-level bureaucrats, public managers, elected officials, and citizens. It also does not address the specific actions government organizations are taking to safeguard accountability as a public value when they use AI systems. Further, when looking at the actual practice of using AI systems in decision-making, state governments are introducing AI legislation to govern the use of technology. City agencies are implementing initiatives to strengthen accountability and oversight in the use of

PROOFS

AI-based systems for public services, which shows that this issue is relevant to both research and practice. However, this does not seem to be enough. In addition to developing ethical governance standards and setting up AI governance boards, and building AI inventories to track whether AI is used and who is responsible for that use (e.g., European Commission, High-Level Expert Group on Artificial Intelligence 2019), scholars and practitioners also encourage the participation of individuals and groups across sectors and disciplines to provide opportunities to include the expertise and experience of all stakeholders. These are, therefore, important areas for future research on AI and public values, with accountability among the most important.

Given the need for further research in these areas, future studies could focus on how AI adoption reshapes accountability relationships between street-level bureaucrats, managers, and algorithm designers. We also know that some systems make decisions autonomously, while others provide recommendations or guidance to government staff, who then make the final decisions. These variations in autonomy and human involvement may influence how citizens perceive accountability. As Chen et al. (2026) argue, more research is, therefore, needed to examine how different types of AI–human collaboration affect public perceptions of accountability. Additional studies could also address what accountability structures and practices are needed when decisions are partially or fully automated. In this respect, one important research question to ask would be, "What mechanisms are currently used by government agencies that help safeguard accountability as a good governance practice that promotes public value creation?" Future research could also study how public servants understand their responsibility when AI recommendations conflict with professional judgment.

4.5 CONCLUSION

The integration of AI into the public sector represents more than a technological upgrade: It is a transformation that reshapes the relationship between governments and their citizens. AI has the potential to improve policymaking, public service delivery, and internal governmental operations, as seen in Section 4.2. Nevertheless, as the examples described in Sections 4.3 and 4.4 demonstrate, the implementation of flawed AI systems can have serious consequences, including upending the livelihoods of citizens, restricting their freedom, or even breaking up their families. As governments move beyond narrow-purpose AI for predictions and classifications and adopt general-purpose generative AI, they are increasingly deploying powerful systems through informal experimentation rather than deliberate institutional design, bypassing the democratic deliberation that ideally legitimates public decisions.

This chapter has identified three interconnected challenges posed by AI in the public sector. First, AI as an optimizing technology conflicts with the multidimensional, normative objectives of public service. The alignment problem

PROOFS

in government is not merely technical but profoundly political, as algorithmic systems surface societal conflicts that democratic institutions have historically managed through "incompletely theorized agreements" (Sunstein 1995). Second, decades of outsourcing have created deep information and power asymmetries between governments and private vendors. The contracting state risks ceding control to private vendors over service delivery, internal operations, and policymaking. Third, the opacity and diffusion of accountability across different relations of responsibility – vertical, horizontal, and diagonal – require existing frameworks of accountability to adapt. Whereas citizens want to hold elected officials, AI developers, and other decision-makers responsible, formal accountability frameworks do not yet offer full opportunities for doing so. Agents seek to avoid accountability by scapegoating AI. The demand for accountability can thus remain unmet.

These challenges are serious but not insurmountable. As we proposed in Section 4.4, addressing the challenges will require transparency as a foundational prerequisite, sociotechnical methods to align AI systems with public values, sustained efforts to rebuild state capacity, and new accountability institutions. Instead of simply asking if AI can make governmental services and operations more efficient, policymakers and civil servants should also consider whether algorithmic governance can be made compatible with democratic accountability. Political scientists and public administration scholars can contribute to these policy debates by examining the interactions among people, institutions, and AI systems.

APPENDIX

TABLE 4.A *Taxonomy of AI use by local, regional/state, and national governments with examples*

AI use activity	Policymaking	Public service delivery	Internal management
(1) Content Creation: "generating new artifacts such as video, narrative, software code, synthetic data"	Local government in Bellingham, Washington (US), is using ChatGPT to draft policy documents (Sanford 2025).	Local government in Reading, Massachusetts, uses a GenAI system to create most of its public-facing communications, including press releases (Jackley 2024).	UK civil servants in the Cabinet Office, Department for Science, Innovation and Technology (DSIT), and No. 10 use LLM Redbox to draft reports, letters, and memos (UK Government 2024).

(*continued*)

PROOFS

TABLE 4.A (*continued*)

AI use activity	Policymaking	Public service delivery	Internal management
(2) Content Synthesis: "combining and/or summarizing parts, elements, or concepts into a coherent whole"	The Massachusetts Platform for Legislative Engagement uses AI to summarize legislation and tag each bill with relevant topics (Massachusetts Platform for Legislative Engagement 2026).	Social workers in England are using AI to summarize notes from meetings and suggest follow-up actions (Booth 2024a).	LLM Redbox synthesizes lengthy government documents for civil servants in the UK Cabinet Office, No. 10, and Department for Science, Innovation and Technology (UK Government 2024).
(3) Prediction: "forecasting the likelihood of a future outcome"	Paris-Saclay agglomeration of municipalities in France uses AI to simulate different energy management scenarios through a digital twin of their territory. This tool forecasts the environmental and financial impacts of projects to improve long-term planning (OECD 2024a).	US National Oceanic and Atmospheric Administration uses AI in global weather forecasting (National Oceanic and Atmospheric Administration 2024).	US Department of the Treasury is using machine learning to forecast economic trends (US Department of the Treasury 2024).
(4) Detection: "identifying, by careful search, examination, or probing, the existence or presence of [something]"	The US Food and Drug Administration (FDA) has, since 2016, piloted the use of AI in postmarket surveillance to update rulemaking and guidance (Engstrom et al. 2020).	The UK uses an ML program to detect fraud in claims for universal credit payments (Booth 2024b).	The US Securities and Exchange Commission uses ML to uncover potential insider trading and inaccuracies in financial reporting (Bandy et al. 2024).

(*continued*)

PROOFS

TABLE 4.A (*continued*)

AI use activity	Policymaking	Public service delivery	Internal management
(5) Digital Assistance: "acting as a personal agent for understanding and responding to commands and questions, and carrying out requested tasks in a conversational manner"	POPVOX Foundation developed Retrieval-Augmented Generation (RAG) AI assistant targeted at US Congressional staffers drafting legislation (POPVOX Foundation 2024).	Kela (The Social Insurance Institution of Finland) uses a chatbot named Kela-Kelpo (or FPA-Folke in Swedish) to assist customers with social security and benefit-related questions (International Social Security Association 2024).	Canada is piloting CANChat to support government employees with research and drafting tasks (Government of Canada 2024).
(6) Discovery: "finding, recognizing, or unearthing something for the first time"	New Jersey's State AI Task Force used an open-source AI-based toolkit to automate web-based research and generate a list of challenges related to the potential negative impacts of AI on workers, the economy, and the labor market (Reboot Democracy 2024).	*Proof-of-concept*: In Togo, during the COVID-19 pandemic, researchers used machine learning algorithms and data from satellites and mobile phone networks to discover people who would have been otherwise inaccurately excluded from social assistance (J-PAL 2020).	The US Department of Homeland Security uses NLP to analyze and extract significant topics/themes from unstructured text responses to open-ended questions in employee surveys (US Department of Homeland Security 2024).
(7) Image Analysis: "recognizing attributes within digital images to extract meaningful information"	*Proof-of-concept*: Deep learning models can be used to analyze land use in cities and aid policymakers in urban planning (OECD 2024b).	US airports are using AI facial recognition to verify passengers (Chung 2025).	Brazil's DETER system uses AI to analyze satellite imagery to monitor and detect deforestation, fires, and illegal activities in the Amazon (Revista Pesquisa FAPESP 2023).

(*continued*)

PROOFS

TABLE 4.A (*continued*)

AI use activity	Policymaking	Public service delivery	Internal management
(8) Information Retrieval/ Search: "finding information about specific topics of interest"	Chat-EUR-Lex is an AI-powered chatbot interface to make EU laws more accessible and searching using retrieval-augmented generation to analyze complex legal texts (Cherubini et al. 2024).	The Vietnamese government operates an AI-powered platform that provides personalized responses to businesses inquiring about administrative requirements (OpenGov 2025).	US federal agencies are testing AI to locate records within government datasets and suggest redactions of information to respond to Freedom of Information Act requests (Kamb 2023).
(9) Monitoring: "observing, checking, and watching over the process, quality, or state of [something] over time to gain insights into how [something] is behaving or performing"	The US House of Representatives introduced an AI tool to automate the process of analyzing differences between bills, amendments, and current laws (POPVOX 2024).	Several US states, including Arkansas, use electronic visit verification systems to monitor in-home personal care services paid for by Medicaid (Eubanks and Mateescu 2021).	The Chinese government used AI to monitor the behavior of public sector employees to catch embezzlement, abuse of power, and misuse of government funds (Griffin 2023).
(10) Performance Improvement: "improving quality and efficiency of the intended outcomes"	*Proof-of-concept*: Researchers have piloted using large language models to optimize economic policy (Gasztowtt et al. 2024).	Los Angeles County, California using a predictive algorithm to identify those who are most vulnerable to becoming homeless so that the government can offer them assistance (Kendall 2024).	The state of Michigan implemented an AI system to improve the efficiency of the procurement process, including automating price comparisons across proposals, generating solicitation change notices, and performing spend analysis (NASPO 2024).

(*continued*)

PROOFS

TABLE 4.A (*continued*)

AI use activity	Policymaking	Public service delivery	Internal management
(11) Personalization: "designing and tailoring [something] to meet an individual's characteristics, preferences, or behaviors"	Politicians in the US are using AI to personalize text and image-based communication with constituents about policy (Daisychain 2026).	Singapore's LifeSG app recommends personalized services for life events (birth, marriage, unemployment) (GovTech Singapore 2024).	US Citizenship and Immigration Services uses generative AI to personalize training for immigration officer personnel (US Department of Homeland Security 2024).
(12) Recommendation: "selecting a course of action from among possible alternatives in order to arrive at a solution"	*Proof-of-concept*: Researchers using AI to automate regulatory impact assessment of different policies (Universidade Nova de Lisboa 2024).	Allegheny Family Screening Tool (AFST) assists call screeners in Allegheny County, Pennsylvania, in deciding whether to investigate child neglect reports based on an algorithmically generated risk score (Eubanks 2018).	The US Social Security Administration uses a recommendation system (with a risk score) to help with disability claim adjudications (Glaze et al. 2022).
(13) Process Automation: "performing repetitive tasks, removing bottlenecks, reducing errors and loss of data, and increasing efficiency of a process"	San Francisco City Attorney's Office used an LLM developed by Stanford University to identify all legislatively mandated reporting requirements, many of which are burdensome and can serve little purpose after decades (Surani et al. 2025).	Arizona, New Jersey, Ohio, Oklahoma, Texas, and Washington are piloting a new AI program to determine whether certain care is eligible for Medicare coverage (Goldstein-Street 2025).	US General Services Administration uses an AI system to automatically review Information and Communication Technology solicitation documents to meet accessibility requirements (US General Services Administration 2024).

Source: The definition of each AI use activity is quoted from Theofanos, Choong, and Jensen (2024).

PROOFS

BIBLIOGRAPHY

Ada Lovelace Institute (2020) "Transparency mechanisms for UK public sector algorithmic decision-making systems," Ada Lovelace Institute. Available at: https://tinyurl.com/253uxjpj.

AI Incident Database (2024) "AI incident database." Available at: https://incidentdatabase.ai/.

Aldemir, C. and Uçma Uysal, T. (2024) "AI competencies for internal auditors in the public sector," *EDPACS*, 69(1), pp. 3–21. Available at: https://doi.org/10.1080/07366981.2024.2312001.

Amodei, D. et al. (2016) "Concrete problems in AI safety," arXiv. Available at: https://doi.org/10.48550/arXiv.1606.06565.

Anderson, A. G. and Linowes, D. (1988) *Privatization, toward more effective government: Report of the president's commission on privatization*. Washington, DC: The Commission. Available at: https://catalog.hathitrust.org/Record/001528084.

Bagley, N. (2019) "The procedure fetish," *Michigan Law Review*, 118(3), p. 345–401. Available at: https://doi.org/10.36644/mlr.118.3.procedure.

Bandy, A., Haffner, E. and Suárez, M. (2024) "The US government is using AI to detect potential wrongdoing, and companies should too," *Skadden*. Available at: https://tinyurl.com/2rmr3vcz.

Bar-Siman-Tov, I. (2025) "Legislatures and legislation in the age of artificial intelligence," *The Theory and Practice of Legislation*. Available at: www.tandfonline.com/doi/abs/10.1080/20508840.2025.2590971.

Benbya, H., Davenport, T. H. and Pachidi, S. (2020) "Artificial intelligence in organizations: Current state and future opportunities," *MIS Quarterly Executive*, 19(4). Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3741983.

Bloch-Wehba, H. (2021) "Transparency's AI problem," *Knight First Amendment Institute*. Available at: https://knightcolumbia.org/content/transparencys-ai-problem.

Bommasani, R. et al. (2021) "On the opportunities and risks of foundation models," arXiv. https://doi.org/10.48550/arXiv.2108.07258.

Bommasani, R. et al. (2024) *The foundation model transparency index v1.1 May 2024*. Palo Alto, CA: Stanford University.

Booth, R. (2024a) "Social workers in England begin using AI system to assist their work," *The Guardian*. Available at: www.theguardian.com/society/2024/sep/28/social-workers-england-ai-system-magic-notes.

Booth, R. (2024b) "Revealed: Bias found in AI system used to detect UK benefits fraud," *The Guardian*. Available at: https://tinyurl.com/276e3rj4.

Bovens, M. (2007) "Analysing and assessing accountability: A conceptual framework," *European Law Journal*, 13(4), pp. 447–468. Available at: https://doi.org/10.1111/j.1468-0386.2007.00378.x.

Bovens, M. and Zouridis, S. (2002) "From street-level to system-level bureaucracies: How information and communication technology is transforming administrative discretion and constitutional control," *Public Administration Review*, 62(2), pp. 174–184.

Bozeman, B. (2017) *Public values and public interest: Counterbalancing economic individualism*. Washington, DC: Georgetown University Press (Public management and change series).

PROOFS

Brunjes, B. M. (2020) "Competition and federal contractor performance," *Journal of Public Administration Research and Theory*, 30(2), pp. 202–219.
Bueno de Mesquita, E. (2019) "The perils of quantification," *Boston Review*, March 26. Available at: https://tinyurl.com/2j5ecxab.
Bullock, J., Young, M. M. and Wang, Y. F. (2020) "Artificial intelligence, bureaucratic form, and discretion in public service," *Information Polity*, 25(4), pp. 491–506.
Busuioc, M. (2021) "Accountable artificial intelligence: Holding algorithms to account," *Public Administration Review*, 81(5), pp. 825–836. Available at: https://doi.org/10.1111/puar.13293.
Callahan, D. and T. Davies. (2017) "Judge: IBM owes Indian $78M for failed welfare automation." Associated Press, August 7. Available at: https://apnews.com/general-news-8eb53eb9bdf94adb92e5b8b09559d8d0
Cameron: Callahan, D. and T. Davies. (2017) "Judge: IBM owes Indian $78M for failed welfare automation." Associated Press, August 7. Available at: https://apnews.com/general-news-8eb53eb9bdf94adb92e5b8b09559d8d0
Cameron, D. (2021) "Researchers: Instagram 'bullied' us into halting algorithmic research," *Gizmodo*, August 13. Available at: https://gizmodo.com/researchers-instagram-bullied-us-into-halting-algorith-1847484811.
Casper, S. et al. (2024) "Black-box access is insufficient for rigorous AI audits," in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY: Association for Computing Machinery (FAccT '24), pp. 2254–2272. Available at: https://doi.org/10.1145/3630106.3659037.
Charette, R. (2018) *Michigan's MiDAS unemployment system: Algorithm alchemy created lead, not gold*, IEEE Spectrum. Available at: https://tinyurl.com/35c3f754.
Chen, T., Gascó-Hernandez, M. and Gil-García, R. (2026) "Understanding accountability in AI use within the public sector: Insights from state governments in the United States." *Public Administration* (special issue on Ethical, Responsible, and Value-Oriented Use of Artificial Intelligence in Public Organizations).
Cherubini, M. et al. (2024) "Improving the accessibility of EU laws: The Chat-EUR-Lex project," in *CEUR Workshop Proceedings. Ital-IA 2024: 4th National Conference on Artificial Intelligence*. Available at: https://ceur-ws.org/Vol-3762/545.pdf.
Chouldechova, A. (2017) "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, 5(2), pp. 153–163. Available at: https://doi.org/10.1089/big.2016.0047.
Christensen, T. and Lægreid, P. (2016) "Introduction: Accountability and welfare state reforms," *The Routledge handbook to accountability and welfare state reforms in Europe*. New York, NY: Routledge, pp. 15–26.
Chung, C. (2025) "Flying abroad? Get ready for a biometric 'corridor,'" *The New York Times*, December 8. Available at: www.nytimes.com/2025/12/08/travel/facial-recoginition-flights-airports-orlando.html.
City of San José (2025) *Government AI coalition*. Available at: https://tinyurl.com/2nsxrvn9.
Coase, R. H. (1937) "The nature of the firm," *Economica*, 4(16), pp. 386–405.
Coglianese, C. (2023) "AI, due process, and trade secrets," *The Regulatory Review*. Available at:www.theregreview.org/2023/09/04/coglianese-ai-due-process-and-trade-secrets/.

PROOFS

Coglianese, C. and Crum, C. R. (2025) "Leashes, not guardrails: A management-based approach to artificial intelligence risk regulation," *Risk Analysis*, 45(12), pp. 4397–4407.

Cooper, A. F. et al. (2022) "Accountability in an algorithmic society: Relationality, responsibility, and robustness in machine learning," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY: Association for Computing Machinery (FAccT '22), pp. 864–876. Available at: https://doi.org/10.1145/3531146.3533150.

Cortez, E. K. and Maslej, N. (2023) "Adjudication of artificial intelligence and automated decision-making cases in Europe and the USA," *European Journal of Risk Regulation*, 14(3), pp. 1–19.

Coyle, D. and Weller, A. (2020) "'Explaining' machine learning reveals policy challenges," *Science*, 368(6498), pp. 1433–1434.

Daisychain (2026) *Texting that turns outreach into action*. www.daisychain.app/features/texting.

Digital Agency, Government of Japan (2025) *The guideline for Japanese governments' procurements and utilizations of generative AI for the sake of evolution and innovation of public administration*. Digital Society Promotion Standard Guidelines DS-920. Available at: https://tinyurl.com/4dwrdw5n.

Eckhouse, L. (2022) "Metrics management and bureaucratic accountability: Evidence from policing," *American Journal of Political Science*, 66(2), pp. 385–401. Available at: https://doi.org/10.1111/ajps.12661.

Engstrom, D. F. et al. (2020) "Government by algorithm: Artificial intelligence in federal administrative agencies," NYU School of Law, Public Law Research Paper No. 20–54.

Erkkilä, T. (2007) "Governance and accountability – A shift in conceptualisation," *Public Administration Quarterly*, 31(1/2), pp. 1–38.

Eubanks, V. (2018) *Automating inequality: How high-tech tools profile, police, and punish the poor*. New York, NY: St. Martin's Press.

Eubanks, V. and Mateescu, A. (2021). "'We don't deserve this': New app places US caregivers under digital surveillance," *The Guardian*, July 28. https://tinyurl.com/yyxfdmkw.

European Commission (2025a) *The general-purpose AI code of practice*. European Commission. Available at: https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai.

European Commission (2025b) *Updated EU AI model contractual clauses: Public buyers community*. Available at: https://tinyurl.com/4m9nsfdm.

European Commission, High-Level Expert Group on Artificial Intelligence (2019) *Ethics guidelines for trustworthy AI*. Publications Office of the European Union. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Fabris, A. et al. (2022) "Algorithmic fairness datasets: The story so far," *Data Mining and Knowledge Discovery*, 36(6), pp. 2074–2152. Available at: https://doi.org/10.1007/s10618-022-00854-z.

Fleck, G. (2024) "A Brazilian municipality has approved a bill written by ChatGPT," *Global Voices*, January 22. Available at: https://tinyurl.com/yuxhmhxk.

Fleisher, W., Cibralic, B., Basl, J., Ricks, V. and Smith, M. N. (2025) "Responsibility and accountability in an algorithmic society," *Philosophy & Technology*, 38(4). https://doi.org/10.1007/s13347-025-00970-w.

PROOFS

Garrett, B. L. (2020) "Justice in forensic algorithms," *Harvard Data Science Review*, 2(1). Available at: https://doi.org/10.1162/99608f92.a6b843af.

Gasztowtt, H. et al. (2024) "Large legislative models: Towards efficient AI policymaking in economic simulations," arXiv. Available at: https://doi.org/10.48550/arXiv.2410.08345.

Glaze, K., Ho, D. E., Ray, G. and Tsang, C. (2022) "Artificial intelligence for adjudication: The Social Security Administration and AI governance," in J. B. Bullock, Y.-C. Chen, J. Himmelreich, V. M. Hudson, A. Korinek, M. M. Young and B. Zhang (eds.) *The Oxford handbook of AI governance* (pp. 779–796). Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197579329.013.46.

Global Partnership on Artificial Intelligence (2024) *Algorithmic transparency in the public sector: A state-of-the-art report*. GPAI/OECD. Available at: https://tinyurl.com/3bzcft54.

Goldstein-Street, J. (2025) *AI reviews rolling out for Medicare in Washington state for some procedures, Oregon Public Broadcasting*. Available at: https://tinyurl.com/4mfx6x4v.

Goodall, N. (2019) "More than trolleys: Plausible, ethically ambiguous scenarios likely to be encountered by automated vehicles," *Transfers*, 9(2), pp. 45–58. Available at: https://doi.org/10.3167/TRANS.2019.090204.

Government of Canada (2024) *CANChat: SSC's first generative AI chatbot*. Available at: https://tinyurl.com/2by2jbdv.

GovTech Singapore (2024) *LifeSG: Digital services for citizens*. Available at: www.tech.gov.sg/products-and-services/for-citizens/digital-services/lifesg/.

Green, B. (2022) "The flaws of policies requiring human oversight of government algorithms," *Computer Law & Security Review*, 45, p. 105681. Available at: https://doi.org/10.1016/j.clsr.2022.105681.

Greenstein, S. M. (1993) "Did installed base give an incumbent any (measurable) advantages in federal computer procurement?," *RAND Journal of Economics*, 24(1), pp. 19–39.

Griffin, M. (2023) "China's AI anti-corruption program gets shut down for being too good," *311 Institute*. Available at: https://tinyurl.com/ypyw6j7a.

Hart, O., Shleifer, A. and Vishny, R. W. (1997) "The proper scope of government: Theory and an application to prisons," *The Quarterly Journal of Economics*, 112(4), pp. 1127–1161.

Himmelreich, J. (forthcoming) "Democracy and the AI-augmented administrative state," in A. M. Messina and B. Fernandez Milmanda (eds.) *Democracy's mega challenges: Climate change, migration, and artificial intelligence*. London, UK: Palgrave Macmillan.

Ho, D. E. (2023) "Opportunities and risks of artificial intelligence in the public sector," *Stanford Law School Blogs*. https://tinyurl.com/ynvpzyj6.

Holland, A., and Nair, G. (eds.) (forthcoming) *The contracting state: How and why governments rely on private contractors*. Unpublished manuscript.

Honig, D. (2024) *Mission driven bureaucrats: Empowering people to help government do better*. Oxford: Oxford University Press.

Hood, C. (1995) "The 'new public management' in the 1980s: Variations on a theme," *Accounting, Organizations and Society*, 20, pp. 93–109. https://doi.org/10.1016/0361-3682(93)e0001-w.

PROOFS

Humane Intelligence (2024) *ARIA-NIST red teaming*. Available at: https://humane-intelligence.org/get-involved/events/aria-nist-red-teaming/.

International Social Security Association (2024) *Artificial intelligence in social security institutions: The case of intelligent chatbots*. Available at: www.issa.int/analysis/artificial-intelligence-social-security-institutions-case-intelligent-chatbots.

Jackley, M. (2024) *Using AI in local government: 10 use cases*, *Oracle*. Available at: www.oracle.com/artificial-intelligence/ai-local-government/.

James, O., Jilke, S., Petersen, C. and Van de Walle, S. (2016) "Citizens' blame of politicians for public service failure: Experimental evidence about blame reduction through delegation and contracting," *Public Administration Review*, 76(1), pp. 83–93. https://doi.org/10.1111/puar.12471.

Jeantet, D. and Savarese, M. (2023). *Brazilian city enacts an ordinance that was secretly written by ChatGPT*. Associated Press, November 30. https://tinyurl.com/wrh997ey.

J-PAL (2020) *Using mobile phone and satellite data to target Togo's emergency cash transfer program*. Available at: https://tinyurl.com/ahn72e2c.

Kamarck, E. (2025) *Is government too big? Reflections on the size and composition of today's federal government*. Washington, DC: Brookings Institution. Available at: https://tinyurl.com/zwpspuzs.

Kamb, L. (2023) "Some U.S. government agencies are testing out AI to help fulfill public records requests," *NBC News*, August 1. Available at: www.nbcnews.com/news/us-news/federal-agencies-testing-ai-foia-concerns-rcna97313.

Kapur, D. (2020) "Why does the Indian state both fail and succeed?," *Journal of Economic Perspectives*, 34(1), pp. 31–54.

Kavi, A. (2025) "Judge blocks detention of British researcher who scrutinizes online hate," *The New York Times*, December 25. Available at: www.nytimes.com/2025/12/25/us/politics/imran-ahmed-judge-order.html.

Kayser-Bril, N. (2024) *Election Instagram algorithm analysis, algorithm watch*. Available at: https://algorithmwatch.org/en/election-instagram-algorithm-analysis/.

Kendall, M. (2024, March 6). "This California county is testing AI's ability to prevent homelessness," *CalMatters*. https://tinyurl.com/4c3bspwd.

Kerkhoff, A. D. N. (2016) "Public value dynamics," in A. Farazmand (ed.) *Global encyclopedia of public administration, public policy, and governance*. Cham: Springer, pp. 1–9.

Kettl, D. F. (1993) *Sharing power: Public governance and private markets*. Washington, DC: Brookings Institution Press.

Khera, R. and Moharil, A. (2024) "Aadhaar: Costs of digital red tape," *Economic and Political Weekly*, 59(19). www.epw.in/journal/2024/19/insight/aadhaar-costs-digital-red-tape.html.

Kim, J. Y. and Kesari, A. (2026) "How State Governments Should Purchase AI to Ensure Fair, Transparent, and Accountable Use," Washington, DC: Federation of American Scientists. Available at: https://fas.org/publication/how-governments-purchase-aifair/.

Kleinberg, J., Mullainathan, S. and Raghavan, M. (2016) "Inherent Trade-Offs in the Fair Determination of Risk Scores," arXiv. Available at: https://doi.org/10.48550/arXiv.1609.05807.

Kolkman, D. (2020) "'F**k the algorithm'?: What the world can learn from the UK's A-level grading fiasco – Impact of Social Sciences," *London School of Economics Impact Blog*, August 26. Available at: https://tinyurl.com/m9tcdvnc.

PROOFS

Kosack, S. and Fung, A. (2014) "Does transparency improve governance?," *Annual Review of Political Science*, 17, pp. 65–87. Available at: https://doi.org/10.1146/annurev-polisci-032210-144356.

Larson, J. et al. (2016) "How we analyzed the COMPAS recidivism algorithm," *ProPublica*. Available at: www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Lawrence, C., Cui, I. and Ho, D. E. (2023) "The bureaucratic challenge to AI governance: An empirical assessment of implementation at U.S. Federal Agencies," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY: Association for Computing Machinery (AIES '23), pp. 606–652. Available at: https://doi.org/10.1145/3600211.3604701.

Lechterman, T. M. (2024) "The concept of accountability in AI ethics and governance," in J. B. Bullock, Y.-C. Chen, J. Himmelreich, V. M. Hudson, A. Korinek, M. M. Young and B. Zhang (eds.) *The Oxford handbook of AI governance* pp. 164–182. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197579329.013.10

Lessig, L. (1999) *Code: And other laws of cyberspace*. New York, NY: Basic Books.

Li, J. (2022) "Value-based payments in health care: Evidence from a nationwide randomized experiment in the home health sector," *Journal of Policy Analysis and Management*, 41(4), pp. 1090–1117. Available at: https://doi.org/10.1002/pam.22415.

Light, P. C. (1999) *The true size of government*. Washington, DC: Brookings Institution Press.

Light, P. C. (2017) *The true size of government*. New York, NY: The Volcker Alliance. Available at: www.volckeralliance.org/resources/true-size-government-1.

Lipsky, M. (1980). *Street-level bureaucracy: Dilemmas of the individual in public services*. New York, NY: Russell Sage Foundation.

Liss, S. and Pradhan, R. (2024) "Errors in Deloitte-run Medicaid systems can cost millions and take years to fix," KFF Health News. Available at: https://tinyurl.com/mpk84ne9.

Low, H. and Pistaferri, L. (2025) "Disability insurance: Error rates and gender differences," *Journal of Political Economy*, 133(9), pp. 2962–3018.

Ludwig, J. and Mullainathan, S. (2021) "Fragile algorithms and fallible decision-makers: Lessons from the justice system," *Journal of Economic Perspectives*, 35(4), pp. 71–96.

Lührmann, A., Marquardt, K. L. and Mechkova, V. (2020) "Constraining governments: New indices of vertical, horizontal, and diagonal accountability," *American Political Science Review*, 114(3), pp. 811–820. Available at: https://doi.org/10.1017/S0003055420000222.

Madan, R. and Ashok, M. (2023) "AI adoption and diffusion in public administration: A systematic literature review and future research agenda," *Government Information Quarterly*, 40(1), p. 101774. Available at: https://doi.org/10.1016/j.giq.2022.101774.

Mallaby, S. (2022) *The power law: Venture capital and the making of the new future*. New York, NY: Penguin Press.

Maryland Office of the Governor (2025) *Maryland governor Wes Moore announces landmark AI partnership to transform state service delivery*. Available at: https://tinyurl.com/2km2n9x3.

PROOFS

Maslej, N. et al. (2023) *The AI index 2023 annual report*. Stanford, CA: AI Index Steering Committee, Institute for Human-Centered Artificial Intelligence, Stanford University.

Massachusetts Platform for Legislative Engagement (2026) *How MAPLE uses AI*, *Massachusetts platform for legislative engagement*. Available at: www.mapletestimony.org/about/how-maple-uses-ai.

Mathur, V. and Zimmerman, M. (2024) *2024 federal agency AI use case inventory*. Available at: https://github.com/ombegov/2024-Federal-AI-Use-Case-Inventory.

McKenzie, P. (2010) "Falsehoods programmers believe about names." Available at: www.kalzumeus.com/2010/06/17/falsehoods-programmers-believe-about-names/.

Meier, K. J. (1997) "Bureaucracy and democracy: The case for more bureaucracy and less democracy," *Public Administration Review*, 57(3), p. 193–199. Available at: https://doi.org/10.2307/976648.

Mettler, S. (2011) *The submerged state: How invisible government policies undermine American democracy*. Chicago: University of Chicago Press.

Moore, M. H. (1995) *Creating public value: Strategic management in government*. Cambridge, MA: Harvard University Press.

Morgan, K. J. and Campbell, A. L. (2011) *The delegated welfare state: Medicare, markets, and the governance of social policy*. Oxford: Oxford University Press.

Mulgan, R. (2000) "'Accountability': An ever-expanding concept?," *Public Administration*, 78(3), pp. 555–573.

Myers, S. L. (2025) "Trump administration cancels scores of grants to study online misinformation," *The New York Times*, May 15. Available at: www.nytimes.com/2025/05/15/business/trump-online-misinformation-grants.html.

Nabatchi, T. (2018) "Public values frames in administration and governance," *Perspectives on Public Management and Governance*, 1(1), pp. 59–72. Available at: https://doi.org/10.1093/ppmgov/gvx009.

Nabatchi, T., Goerdel, T. H. and Peffer, S. (2011) "Public administration in dark times: Some questions for the future of the field," *Journal of Public Administration Research and Theory*, 21(Supplement 1), pp. i29–i43. Available at: https://doi.org/10.1093/jopart/muq068.

NASPO (2024) *AI sets sail: Early waves in state procurement innovation*. Available at: www.naspo.org/news/ai-sets-sail-early-waves-in-state-procurement-innovation/.

National Internet Observatory, Northeastern University (2024) *National internet observatory*. Available at: https://nationalinternetobservatory.org.

National Oceanic and Atmospheric Administration (2024) *NOAA deploys new generation of AI-driven global weather models*. Available at: https://tinyurl.com/39hvfyvx.

Novelli, C., Taddeo, M. and Floridi, L. (2023) "Accountability in artificial intelligence: What it is and how it works," *AI & Society*, 39, pp. 1871–1882.

Obermeyer, Z. et al. (2019) "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, 366(6464), pp. 447–453. Available at: https://doi.org/10.1126/science.aax2342.

O'Donnell, G. A. (1998) "Horizontal accountability in new democracies," *Journal of Democracy*, 9(3), pp. 112–126. Available at: https://doi.org/10.1353/jod.1998.0051.

PROOFS

OECD (2024a) *Governing with artificial intelligence: Are governments ready?* OECD Artificial Intelligence Papers 20. Paris: OECD Publishing. Available at: https://doi.org/10.1787/26324bc2-en.
OECD (2024b) *Monitoring land use in cities using satellite imagery and deep learning*. Available at: https://tinyurl.com/praap9fp.
OECD (2025) *Government at a glance 2025*. OECD Publishing (Government at a Glance). https://doi.org/10.1787/oefdobcd-en.
Office of Management and Budget (1999) *Circular no. A-76*. Available at: https://tinyurl.com/yt4bjb6m.
OpenGov (2025) *Vietnam: AI-driven solutions for smarter public services – OpenGov Asia, OpenGov*. Available at: https://opengovasia.com/vietnam-ai-driven-solutions-for-smarter-public-services/.
Pahlka, J. (2023) *Recoding America: Why government is failing in the digital age and how we can do better*. New York, NY: WW Norton & Company.
Papadakis, T. et al. (2024) "Explainable and transparent artificial intelligence for public policymaking," *Data & Policy*, 6, p. e10. Available at: https://doi.org/10.1017/dap.2024.3.
POPVOX (2024) *Comparative print suite*. Available at: www.popvox.org/legitech/comparative-print-suite.
POPVOX Foundation (2024) *StaffLink: RAG AI assistant for congressional staffers*. Available at: www.popvox.org/blog/stafflink.
Reboot Democracy (2024) *New Jersey AI task force policy recommendations*. Available at: https://rebootdemocracy.ai/blog/new-jersey-ai-task-force-policy-recommendations.
Revista Pesquisa FAPESP (2023) "Multiple systems use satellites to monitor deforestation in the Amazon," *Revista Pesquisa FAPESP*. Available at: https://tinyurl.com/mum5jm4y.
Sanders, N. and Schneier, B. (2025) "AI will write complex laws," *Lawfare*. Available at: www.lawfaremedia.org/article/ai-will-write-complex-laws.
Sanford, N. (2025). "Washington city officials are using ChatGPT for government work," *Cascade PBS*, August 26. https://tinyurl.com/mwvj99v3.
Schiff, D. S., Schiff, K. J. and Pierson, P. (2021) "Assessing public value failure in government adoption of artificial intelligence," *Public Administration*, 100(3), pp. 653–673. Available at: https://doi.org/10.1111/padm.12742.
Simon, H. A. (1947) *Administrative behavior: A study of decision-making processes in administrative organization*. New York, NY: Macmillan.
Sunstein, C. R. (1995) "Incompletely theorized agreements," *Harvard Law Review*, 108(7), pp. 1733–1772.
Surani, F. *et al*. (2025) *Cleaning up policy sludge: An AI statutory research system*. Palo Alto, CA, USA: Stanford University Human-Centered Artificial Intelligence.
Theofanos, M. F., Choong, Y., and Jensen, T (2024) *AI use taxonomy: A human-centered approach*. NIST 200-1. Gaithersburg, MD: National Institute of Standards and Technology (US), p. NIST 200-1. Available at: https://doi.org/10.6028/NIST.AI.200-1.
Turing, A. M. (1936) "On computable numbers, with an application to the Entscheidungs problem," *Proceedings of the London Mathematical Society*, 43(2), pp. 230–265.
US Government Accountability Office (2023) *Artificial intelligence: Key practices to help ensure accountability in federal use*. U. S. Government Accountability Office. Available at: www.gao.gov/products/gao-23-106811.

PROOFS

US Government Accountability Office (2025) *Artificial intelligence: Generative AI use and management at federal agencies*. Report to Congressional Requesters. United States Government Accountability Office. Available at: www.gao.gov/products/gao-25-107653.

UK Government (2024) *DSIT Redbox: Algorithmic transparency record*. Available at: www.gov.uk/algorithmic-transparency-records/dsit-redbox.

UK National Audit Office (2025) *Use of artificial intelligence in government*. U.K. National Audit Office. Available at: https://www.nao.org.uk/reports/use-of-artificial-intelligence-in-government/.

Universidade Nova de Lisboa (2024) *Transforming legislative impact assessment using AI to produce better regulations*. Available at: https://tinyurl.com/mm2v8ha9.

US Department of Homeland Security (2024) *AI inventory*. Available at: www.dhs.gov/archive/data/AI_inventory.

US Department of the Treasury (2024) *US Department of the Treasury's AI strategy for OMB memorandum M-25-21*. US Department of the Treasury. Available at: https://home.treasury.gov/system/files/136/Treasury-AI-Strategy.pdf.

US Food and Drug Administration (2025) *FDA launches agency-wide AI tool to optimize performance for the American people*. Available at: https://tinyurl.com/fdt6hv88.

US General Services Administration (2024) *Solicitation review tool*. Available at: www.section508.gov/buy/solicitation-review-tool/.

US General Services Administration (2025) *GSA, Google announce transformative "Gemini for government" OneGov agreement*. Available at: https://tinyurl.com/yx4hfw4x.

Van Noordt, C. and Misuraca, G. (2022) "Exploratory insights on artificial intelligence for government in Europe," *Social Science Computer Review*, 40(2), pp. 426–444.

Wachter, S., Mittelstadt, B. and Floridi, L. (2017) "Why a right to explanation of automated decision-making does not exist in the general data protection regulation," *International Data Privacy Law*, 7(2), pp. 76–99. Available at: https://doi.org/10.1093/idpl/ipx005.

Wallis, N. (2022) *The great Post Office scandal: The fight to expose a multimillion pound IT disaster that put innocent people in jail*. Bath, UK: Bath Publishing.

Whitford, E. (2021) "Facebook disables NYU research accounts," *Inside Higher Ed*, August 6. Available at: https://tinyurl.com/35wjaac8.

Williamson, O. E. (1981) "The economics of organization: The transaction cost approach," *American Journal of Sociology*, 87(3), pp. 548–577.

Wirtz, B. W., Weyerer, J. C. and Geyer, C. (2019) "Artificial intelligence and the public sector – Applications and challenges," *International Journal of Public Administration*, 42(7), pp. 596–615. Available at: https://doi.org/10.1080/01900692.2018.1498103.

Wu, T. (2018) *The curse of bigness: Antitrust in the new gilded age*. New York, NY: Columbia Global Reports.

Yager, S. (2025) "How the most remote community in America gets its mail," *The Atlantic*, May 7. Available at: www.theatlantic.com/magazine/archive/2025/06/mule-mail-delivery-supai-arizona/682619/.

Young, M. M., Bullock, J. B. and Lecy, J. D. (2019) "Artificial discretion as a tool of governance: A framework for understanding the impact of artificial intelligence on public administration," *Perspectives on Public Management and Governance*, 2(4), pp. 301–313. Available at: https://doi.org/10.1093/ppmgov/gvz014.

PROOFS

Young, M. M. et al. (2021) "Artificial intelligence and administrative evil," *Perspectives on Public Management and Governance*, 4(3), pp. 244–258. Available at: https://doi.org/10.1093/ppmgov/gvab006.

Zouridis, S., van Eck, M. and Bovens, M. (2020) "Automated discretion," in T. Evans and P. Hupe (eds.) *Discretion and the quest for controlled freedom*. Cham, Switzerland: Palgrave Macmillan, pp. 313–329.

PROOFS

5

AI, Race, and Politics

Rachel Gillum, Gregory Leslie, and Cara Wong

PROOFS

5.1 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

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.

Tucker_9781047774734_C005.indd 152 08-Sep-26 2:04:02 PM

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

PROOFS

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.

5.2 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.

5.2.1 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.

PROOFS

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. Large language models 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. Large language models require broad, heterogeneous data to achieve generalizable performance, creating a trade-off 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 trade-off. 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.

5.2.2 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.

PROOFS

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 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.

5.2.3 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 trade-offs regarding which populations bear the costs of algorithmic mistakes.

PROOFS

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 LLMs 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. In the rest of the chapter, we 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.

5.3 LINKING AI TO POLITICAL OUTCOMES

5.3.1 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

PROOFS

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).

5.3.1.1 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, 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., 2023). 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).

PROOFS

Administrative burden theory shows how procedural design structures learning, compliance, and psychological costs, with particularly strong effects for marginalized populations (Ray, Herd, and Moynihan, 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.

5.3.1.2 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., 2012; 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,

PROOFS

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).

5.3.1.3 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.

5.3.2 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,

PROOFS

2024; Jun et al., 2022; Flores and Coppock, 2018). This gap presents a significant research opportunity. Section 5.3.2.1 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 on Information Ecosystem).

5.3.2.1 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. Artificial intelligence–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. Artificial intelligence 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

PROOFS

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)?

5.3.2.2 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, 2021). 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

PROOFS

stronger effects when messengers share ethnic or cultural characteristics with voters (Garcia-Bedolla and Michelson, 2012; Sinclair et al., 2013; Green and Gerber, 2019). Artificial intelligence 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 trade-off 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).

5.3.2.3 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

PROOFS

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). Artificial intelligence–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.

Artificial intelligence–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 compliance with section 5 of the Voting Rights Act) is a political decision, and AI tools amplify the consequences of that choice across millions of voters simultaneously.

5.3.2.4 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

PROOFS

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.

5.4 AI AS A METHODOLOGICAL ACTOR IN POLITICAL SCIENCE

Alongside major consequences for real-world political behavior, AI systems are reshaping political science research itself. Artificial intelligence 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 5.2 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. Chapter 10 in this volume documents

PROOFS

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). 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.

Sections 5.4.1–5.4.8 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.

5.4.1 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).

PROOFS

5.4.2 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.

5.4.3 Translation and Multilingual Research

The measurement validity problems described earlier 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

PROOFS

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, and acknowledging interpretive choices – machine translation requires similarly careful examination.

5.4.4 Estimating the Causal Effects of Race

Artificial intelligence 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, and 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.

PROOFS

The more immediate risk is that algorithmic inference contaminates the identification strategies researchers use. Artificial intelligence'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) developed design-based methods for downstream inference that adjust for imperfect AI surrogates, allowing researchers to account for nonuniform 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, particularly in domains where identity is central to theory and inference.

5.4.5 AI-Mediated Data Collection and Political Expression

Sections 5.4.1–5.4.4 addressed 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. Artificial intelligence 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; Garcia and Stout, 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, 2025; Schuetzler et al., 2018; Ho et al., 2018; Xiao et al., 2020). Recent political science-adjacent work evaluates LLMs 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,

PROOFS

2014). Latino communities in perceived anti-immigrant climates show reduced institutional engagement broadly, including those who are US-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 nonresponse that is systematic and racially correlated rather than random. Artificial intelligence–administered surveys thus risk compounding the very underrepresentation they purport to study.

For studies involving diverse populations, this introduces underexamined measurement concerns. Artificial intelligence–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, and 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.

5.4.6 AI-Mediated Sampling and Population Bias

Sections 5.4.1–5.4.5 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). Artificial intelligence–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

PROOFS

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.

5.4.7 Algorithmic Distortion of Benchmarking Infrastructure

Sections 5.4.1–5.4.6 documented 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.

5.4.8 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., 2025; Bender et al., 2021; Weidinger et al., 2021; Grimmer et al., 2022). By design, LLMs 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

PROOFS

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.

5.5 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 5.4 – 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.

5.5.1 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.

5.5.1.1 Core AIMS Components

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

PROOFS

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

PROOFS

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.

5.5.2 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. Artificial Intelligence Measurement Statements are most valuable precisely in cases where researchers have not yet considered whether their AI tools perform differently across the populations they study.

PROOFS

Example 1: Text Classification in Comparative Research

(1) **Instrument:** We used a multilingual LLM (GPT-4o, 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 the open-source DeepFace library (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 library's built-in demographic classification categories based on facial features and skin tone to assign perceived racial categories. These categories, thresholds, and confidence cutoffs were adopted from the default system configuration. The code is open source, but the training data behind its attribute models is not fully documented, so the features driving racial categorization cannot be fully evaluated.

(3) **Differential Performance:** Prior audits of comparable facial analysis systems 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.

PROOFS

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 interview was being administered by an AI system. 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 NLP, probing logic, and response interpretation are likely to perform unevenly across the three languages. English-language capabilities in GPT-4o are known to be 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 interview was administered by an AI system 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.

PROOFS

5.5.3 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 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 (APSR, n.d.). 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.

Table 5.1 maps each AIMS question to analogous practices in both traditions. Specifically, column 1 of the table is the AIMS question (pertaining to the Instrument, Configuration, Differential Performance, and Validation); column 2 is the Purpose for inference; column 3 is the Analog in political science methods; and column 4 is the Analog in industry practice.

The most directly comparable framework is the Guidelines for Using Interactions in Developmental and Experimental research with Large Language Models (GUIDE-LLM) checklist developed by Feuerriegel et al. (2026) for LLM use in behavioral and social science research. Guidelines for

PROOFS

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 unobservable about how it processed key concepts?	Defines the measurement rule and enables evaluation of construct validity.	Question wording; codebook definitions; coder training materials; documentation of missing or unavailable design information.	Prompt templates and labeling protocols; Model card "Training Data" and "Factors" 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.

PROOF

Using Interactions in Developmental and Experimental research with Large Language Models include 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, the American Political Science Association 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.

5.6 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

PROOFS

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 5.4 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

PROOFS

Tucker_9781047774734_C005.indd 180 08-Sep-26 2:04:03 PM

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.

APPENDIX: AI MEASUREMENT STATEMENT 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,

PROOFS

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.

REFERENCES

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.

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

APSR. n.d. "Preparing your materials." Accessed July 12, 2026. https://tinyurl.com/fntefncj.

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. 2020. "Latinos' Deportation Fears by Citizenship and Legal Status, 2007 to 2018." *Proceedings of the National Academy of Sciences* 117 (16): 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): 1–62.

PROOFS

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

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. 2015. "Understanding the Political Representativeness of Twitter Users." *Social Science Computer Review* 33 (6): 712–729. https://doi.org/10.1177/0894439314558836.

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

Barrie, Christopher, Alexis Palmer, and Arthur Spirling. 2025. "Replication for Language Models: Problems, Principles, and Best Practice for Political Science." *Working Paper*. https://arthurspirling.org/documents/BarriePalmerSpirling_TrustMeBro.pdf.

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.

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.

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*. FAccT '21, March 3–10, 2021, Virtual Event, Canada, 610–623. https://doi.org/10.1145/3442188.3445922.

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*. ACM (Association for Computing Machinery), 289–298. https://dl.acm.org/doi/10.1145/3287560.3287575.

Besco, Randy. 2024. "Discrimination and Political Engagement: A Cross-national Test." *Journal of Race, Ethnicity, and Politics* 9 (3): 472–487.

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. https://arxiv.org/abs/1707.00061.

PROOFS

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. https://proceedings.mlr.press/v81/buolamwini18a.html.

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.

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.

PROOFS

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, www.politico.com/news/2023/10/16/nyc-adams-ai-languages-00121744.
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.
Egami, Naoki, Musashi Hinck, Brandon M. Stewart, and Hanying Wei. 2023. "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.
Egami, Naoki, Musashi Hinck, Brandon M. Stewart, and Hanying Wei. 2024. "Using Large Language Model Annotations for the Social Sciences: A General Framework of Using Predicted Variables in Downstream Analyses." *Working Paper.*
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.

PROOFS

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*. 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.

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.

Freelon, Deen, Charlton D. McIlwain, and Meredith Clark. 2016. "Beyond the Hashtags: # Ferguson, # Blacklivesmatter, and the Online Struggle for Offline Justice." *Center for Media & Social Impact, American University*.

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.

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

PROOFS

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

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.

Harris, Adam S., and Michael G. Findley. 2014. "Is ethnicity identifiable? Lessons from an Experiment in South Africa." *Journal of Conflict Resolution* 58 (1): 4–33.

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)*. Berlin, Germany. Association for Computational Linguistics.

Huber, Gregory A., Marc Meredith, Michael Morse, and Katie Steele. 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.

PROOFS

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. FAccT '21*, March 3–10, 2021, Virtual Event, Canada, 375–385. https://doi.org/10.1145/3442188.3445901.

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. https://tinyurl.com/yrfwumju.

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). https://doi.org/10.1126/sciadv.adl2524.

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

PROOFS

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.

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.

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. 2010. "Political Consequences of the Carceral State." *American Political Science Review* 104 (4): 817–833.

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.

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 *Proceedings of the 21st International Conference on Web Information*

PROOFS

Systems and Technologies (WEBIST 2025). SCITEPRESS – Science and Technology Publications, Lda, 71–82. https://doi.org/10.5220/0013706600003985.
Meng, Amanda and Carl DiSalvo. 2018. "Grassroots Resource Mobilization through Counter-Data Action." *Big Data & Society* 5 (2): 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://tinyurl.com/5fzrz3v9.
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.
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 *FAT* '19: Conference on Fairness, Accountability, and Transparency*, January 29–31, 2019, Atlanta, GA, USA, 220–229. https://doi.org/10.1145/3287560.3287596; arXiv:1810.03993v2.
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. 2019. "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries." *Frontiers in Big Data* 2: 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.

PROOFS

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.
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. 2022. "Understanding Chilling Effects." *Minnesota Law Review* 106 (3): 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 (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.
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 *AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 429–435. https://doi.org/10.1145/3306618.3314244.
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.

PROOFS

Rosa, Jonathan and Nelson Flores. 2017. "Unsettling Race and Language: Toward a Raciolinguistic Perspective." *Language in Society* 46 (5): 621–647.

Ross II, Bertrall L. and Spencer, Douglas M. 2021. "Voter Data, Democratic Inequality, and the Risk of Political Violence." *Cornell Law Review* 107: 1011.

Ruths, Derek, and Jürgen Pfeffer. 2014. "Social Media for Large Studies of Behavior." *Science* 346 (6213): 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.

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

Timmons, Adela C., Jacqueline B. Duong, Natalia Simo Fiallo, Theodore Lee, Huong Phuc Quynh Vo, Matthew W. Ahle, Jonathan S. Comer, LaPrincess C. Brewer, Stacy L. Frazier, and Theodora Chaspari. 2023. "A Call to Action on Assessing and Mitigating Bias in Artificial Intelligence Applications for Mental Health." *Perspectives on Psychological Science* 18 (5): 1062–1096. https://doi.org/10.1177/17456916221134490.

Tyler, Tom R. 1990. *Why People Obey the Law*. Yale University Press.

PROOFS

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. Wiley.

Vargas, Edward D., Gabriel R. Sanchez, and Melina Juárez. 2017. "Fear by Association: Perceptions of Anti-Immigrant Policy and Health Outcomes." *Journal of Health Politics, Policy and Law* 42 (3): 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. 2022. "Artificial Intelligence and the Conduct of Literature Reviews." *Journal of Information Technology* 37 (2): 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.

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.

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

Wu, Cary, Rima Wilkes, and David C. Wilson. 2022. "Race & Political Trust: Justice as a Unifying Influence on Political Trust." *Daedalus* 151 (4): 177–199.

Wuttke, Alexander, Matthias Aßenmacher, Christopher Klamm, Max M. Lang, Quirin Würschinger and Frauke 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)*. Albuquerque, New Mexico, 179–204. https://aclanthology.org/2025.latechclfl-1.17/.

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

PROOFS

Chatbot to Conduct Conversational Surveys with Open-Ended Questions." *ACM Transactions on Computer-Human Interaction (TOCHI)* 27 (3): 1–37.

Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell. 2018. "Mitigating Unwanted Biases with Adversarial Learning." In *AIES'18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM Digital Library, 335–340. https://doi.org/10.1145/3278721.3278779.

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.

PROOFS

6

AI, Gender, and Politics

Dawn Teele, Shira Pindyck, and Sophia Lipkin

PROOFS

6.1 INTRODUCTION

The study of gender and politics, which has normative and empirical dimensions, asks how multi-level processes shape institutions, social relations, and individual behavior, producing access to rights, representation, and power across genders (Boussalis et al. 2021; Cohen and Karim 2022). Many scholars think of gender itself as an institution that is shaped and reproduced by culture, economic systems, and politics. By changing the nature of labor, social life, and culture, artificial intelligence (AI) affects political and economic power (Farrell and Fourcade 2023). Gender dynamics are rooted in these power dynamics. Although there are many definitions and manifestations of power, it is useful to conceive of two forms: "power to" and "power over" (Lazar 2022). Artificial intelligence stands to impact both: AI systems for data creation and analysis have the power to exacerbate gender-based inequality, and dominant groups can use AI to increase their power over others.

This chapter outlines three applications of AI to the study of gender and politics: (1) a tool for data collection; (2) a method of data analysis; and (3) a category of analysis, either as a dependent or independent variable. As a tool for data collection and method of analysis, AI has some potential to both further entrench or upend existing gendered power structures and social biases. Artificial intelligence enables gender scholars to reveal inequality in the world around them and, potentially, level the uneven playing field that has hitherto

All author(s) contributed equally. Our thanks to Divij Sinha, Lisa Keen, Nate Persily, Henry Farrell. We also appreciate the work of the APSA task force on AI, and the leadership of Nate Persily and editorial refinements by Lisa Keen. The views expressed here are those of the authors alone and do not represent the views, policies, or positions of the U.S. Department of War or its components, to include the Department of the Navy or the U.S. Naval War College.

Tucker_9781047774734_C006.indd 195 15-09-2026 07:03:38

existed because of a history of exclusion of women from technical fields and careers. However, because most of the researchers at the top of this field are men, the largest gains from these new tools threaten to accrue to them. As a category of analysis, gender scholars must consider how AI will shape dominant groups' abilities to exert power over others. The bulk of this chapter, therefore, lays out a research agenda for studying AI's impact on gender and power.

6.2 AI AS A TOOL FOR DATA COLLECTION

Applications of AI abound in everyday life. For example, it is being used in generating and analyzing documents by academics, lawyers, and consultants, among others; in individually targeted campaigns or corporate advertisements; as a consumer-support resource; in image-based healthcare diagnostic support; and in algorithms promoting press coverage or consumer products. By early 2026, AI tools have already been seamlessly integrated into the daily lives of people around the world – especially in the richest countries. Yet scholars are still grappling with how best to harness its potential as a tool of research.

Artificial intelligence algorithms can be written to "web-scrape," that is, extract information from websites or retrieve data from online databases, digital platforms, and social media outlets. Artificial intelligence systems can aggregate and collate information from multiple online sources, such as RSS feeds, newspapers, and journal databases. Artificial intelligence–powered Natural Language Processing (NLP) techniques can be used for data extraction, classification, and translation from sources such as articles, blogs, comments, historical documents, and government correspondence. Gender scholars already use text-as-data to study gender dynamics in politics – for example, in the manifestos of political parties (Meguid et al. 2025); in the patterns of interruption during political speeches in Congress (Miller and Sutherland 2023); and in the responses of officeholders to constituents (Teele, Kalla, and Rosenbluth 2018). Recent scholarship has used audio recordings of congressional floor speeches to compare "emotional intensity" – "small changes in vocal pitch that are difficult for speakers to control" – across sexes (Dietrich, Hayes, and O'Brien 2019).

Early adoption of AI to collect political data has shown promise in expanding the breadth of scholarship. Boussalis et al. (2021) use machine learning and computer vision tools during televised debates to extract the emotional expressions of male and female political candidates and the reactions of audience members. Studying more than 595,000 observations from German leadership debates, they found that then–German Chancellor Angela Merkel was more likely to express happiness than her interlocutors, and viewers rewarded her for this emotional display while punishing others for nonemotive expressions. These results speak to decades of scholarship in the study of political behavior, showing how candidates adapt strategically to give voters what they want (Boussalis et al. 2021: 1255). We urge gender scholars to utilize AI techniques as a tool for data creation but note that they must take seriously the evidence

PROOFS

that AI can be inconsistent and unreliable in the realm of gender-based classification (Schwemmer et al. 2020),[1] particularly given that the data utilized and compiled by AI systems can be riddled with bias.

6.3 AI AS A TOOL OF ANALYSIS

Artificial intelligence systems can help scholars identify patterns in large datasets, recognize and classify images and text, and troubleshoot code for statistical analyses and data visualization. However, the use of AI as a tool of analysis for research on gender also runs into the same inherent challenges that produce discrimination and reinforce existing prejudices.[2] In other words, AI systems carry socially constructed expectations regarding the behaviors and characteristics of individuals on the basis of their sex. Such expectations not only produce social biases but can also dampen the analytical power of scholarly research.

Further, there exists an increasingly widening gender gap in usage of AI among researchers: Male scholars use AI more frequently than their female counterparts, and experience "higher efficiency improvement from its use" – widening the gap in research productivity (Tang et al. 2025; Otis et al. 2024). On the other hand, female scholars, can take advantage of AI tools to improve research processes, such as integrating AI tools to assist with coding tasks or web search and scraping, and with writing tasks, improving efficiency, decreasing the costs and time required for large data research, and, at its best, allowing them to be more ambitious in their intellectual projects.

Research on gender and politics can take advantage of AI tools such as NLP to analyze text-based data like speeches and social media content. Natural Language Processing can be used in structured or unstructured ways to engage in sentiment analysis and topic modeling. For example, Funk et al. (2022) used supervised tree-based learning to study the drivers of public spending on education, healthcare, and defense. They found that "education and healthcare spending increases sharply after women's representation in the legislature surpasses 20 and 15 percent, respectively, and then flattens after women's representation reaches approximately 41 percent for education spending and 35 percent for healthcare spending." Machine learning thus allowed them to argue that women's representation matters over an interval, instead of at a discrete inflection point, adding nuance to the classic "critical mass" theory.

[1] Gender classification is more accurate in men than women, with highest errors for Asian women over 50 (Ngan and Grother 2015, i). Artificial intelligence systems are increasingly grappling with gender inferences based on physical attributes. Given these challenges, Google removed gender from its image recognition AI algorithm in 2020 (Aggarwakl et al. 2022; Schwemmer et al. 2020).

[2] Sjoberg et al. (2018); Cohen and Karim (2022).

PROOFS

6.4 AI AS A CATEGORY OF ANALYSIS

Much interest in AI stems from its "revolutionary" promise – its potential to transform wealth, welfare, and power structures for the good. However, AI systems also have the potential to reinforce or exacerbate existing inequalities. As a category of analysis, AI can become the dependent variable – where we study the systems themselves and how gender dynamics and hierarchies impact the development or output of these systems – or AI can be studied as an independent variable – where we study how the system impacts other outcomes of interest.

6.5 AI AS A DEPENDENT VARIABLE

Artificial intelligence algorithms create categories and systematically apply them to new sources of information (Farrell and Fourcade 2023). Typically, the classifications made by AI systems are not visible to the public, yet in many domains of life – credit allocation, teacher tenure, sentencing length by judges, targeted media exposure, and job placement – AI's classification schemes are being used to distribute resources and opportunities. A lack of transparency into AI's classification schemas should be of interest to scholars of gender and politics.

Sociologist Cecilia Ridgeway characterizes gender as a "primary frame" – a simplified cultural-category system that acts as common knowledge – which enables people to classify and interact meaningfully with one another. Given that male–female distinctions are almost universally recognized across societies (albeit with varying degrees of hierarchical ordering), gender binaries have become a routine tool for AI systems tasked with categorizing individuals. This reliance on gender-based distinctions can enhance systemic efficiency, which AI algorithms prioritize. As Farrell and Fourcade (2023) note, algorithms "extend *both* the logic of hierarchy *and* the logic of competition" by using classifications to understand vertical relations between categories and by seeking the fastest ways to make these classifications.

Scholars have begun to study AI as a dependent variable through the lens of "algorithmic bias," by examining how AI choices can further entrench racial and gender-based biases (see Chapter 5 on race). Often, AI systems rely on massive amounts of data to find patterns and make predictions, informing who health-care systems prioritize, which candidates are called for job interviews, and how much credit customers are offered by financial institutions. This data carries biases which are embedded in the geometry of word associations within large language models (LLMs) (Bolukbasi et al. 2016). By encoding "wrongfully discriminatory social patterns into an algorithm," AI systems can support similarly discriminatory social practices (Vredenburgh 2022, 129). Gender bias in these systems can impact psychological health and economic outcomes, and it can amplify prejudices (Smith and Rustagi 2021). Facial recognition systems, for example, are less accurate for women with darker skin (Boussalis et al. 2021). In healthcare, the downstream effects of biased AI systems contribute to gendered disparities in diagnoses (Aggarwal et al. 2022). The design

PROOF

of biomedical AI algorithms often fail to account for sex and gender, producing not only mistakes but also discriminatory outcomes (Cirillo et al. 2020).[3]

Arguably, algorithms are biased because they are built by humans and learn from large corpuses of information that humans have created (O'Neil 2016). The people who develop AI systems and determine what datasets, rules, proxies, and variables are fed into algorithms play a crucial role in determining whether or not a given algorithm will penalize certain communities or identities (Smith and Rustagi 2021). These biases may result from a lack of diversity in the AI industry or from blind spots of AI researchers preoccupied with improving the performance of models without rectifying the underrepresentation of marginalized populations in the training data.[4] Put simply, the AI workforce is mostly male (Mundy 2017), so the processes that contribute to biased representations are not only algorithmic but also social. Large language models, such as ChatGPT, often "parrot" the hegemonic worldviews of their training data, propagating, proliferating, and amplifying discriminatory content (Bender et al. 2021) and require manual intervention to fix (Bolukbasi et al. 2016). In other words, the content individuals view and choose to disseminate contributes to and is further entrenched by biased AI. Gender scholars should take up the question of how algorithmic bias can be detected and measured. Scholars could look under the hood by studying algorithmic inputs or, at the point of contact – by, for example, treating LLMs as the site of research and studying how LLMs respond to classic questions about gender and politics (see also Chapters 5 and 10 on race and methods). The norms, laws, and institutions that shape how AI is built and utilized can impact how algorithmic biases manifest.

6.6 AI AS AN INDEPENDENT VARIABLE

Beyond studying the algorithms themselves, we urge gender scholars to conduct research on how the existence of AI tools, the classification and allocation systems of AI, and the labor-substituting potential of AI impact gender dynamics in institutions and society more generally. A list of potential research questions might include the following:

(1) **How does algorithmic bias impact access to resources and opportunities for men and women?**

Artificial intelligence has the potential to reduce discriminatory bias. In healthcare, for example, when AI algorithms integrate sex and gender differences effectively, they can enhance our understanding of the differentiation of vulnerabilities for disease among individuals (Cirillo et al. 2020). However, the use of LLMs in health service can also reproduce bias and result in gender-based

[3] Genome-wide association studies (GWAS), for example, tend to rely on white male subjects, overlooking potential sex differences in disease. Whose genomics? (2019).

[4] Buolamwini and Gebru (2018).

PROOFS

disparities in service receipt (Rickman 2025). In many domains, algorithms seem to reinforce discrimination against women and minorities.

For example, in the 1970s, St. George's Hospital Medical School's automated assessment program replicated discriminatory policies against applicants on the basis of race and gender (O'Neil 2016, 116–117). In 2019, Apple Card algorithms granted lower credit to women than men, triggering investigations by the US Senate Committee on Finance and the New York State Department of Financial Services (Crawford et al. 2019, 12).[5] In another case, an online tech hiring platform used social data generated by actions in the digital realm (such as time spent sharing and developing community online) to measure candidate participation in the digital community, failing to account for (1) disparities between the time men and women have to chat online (often due to societal expectations regarding unpaid care for children), and (2) women who adopt male identities online to address gender-specific safety concerns and bias. As a result, the platform created an algorithm that systematically ranked female candidates lower than their male counterparts (Smith and Rustagi 2021). Women and ethnic minorities are often excluded from targeted advertisements for rental listings that rely on personalized algorithms (Spinks 2020). Moreover, the initial salary negotiation offers from LLMs are substantively lower for women, people of color, and refugees (Sorokovikova et al. 2025).

Algorithmic bias can also perpetuate racial and gender stereotypes. For example, the voices ascribed to AI caregiving devices providing patients with medication reminders, social companionship, and mental health services are often female (Littman et al. 2021, 68). Smith and Rustagi (2021) studied bias across industries using 133 publicly available algorithmic selection procedures from 1988 to present. They identified six domains of algorithmic bias: (1) lower quality of service, (2) unfair allocation of resources, information, and opportunities, (3) reinforcement of existing stereotypes and prejudices, (4) derogatory and offensive treatment, (5) detriments to physical safety, and (6) health hazards. Each of these arenas suggest further research is needed by gender scholars, especially those who study how bureaucracies and street-level officials interact with and allocate resources in their communities.

(2) How do AI algorithms impact political behavior and representation?

Political behavior is a broad concept that comprises an individual's engagement with politics. Certain areas of political behavior – such as volunteering for campaigns, donating money, learning and talking about politics, and voting – are related to the idea that people have confidence that they can get what they need out of politics. However, AI algorithms can undermine confidence in political processes, especially among marginalized groups. The increasing

[5] For a greater discussion, see Bowdish (2012). Such biases shape women's economic rights. In the United States, marital status and gender were used to determine creditworthiness until the mid 1970s, resulting in widespread credit discrimination.

PROOFS

adoption of algorithms *by* political actors, or their broad use across society, represent "what Ridgeway (2011, 40) termed an 'amplification process,' that is, a mechanism through which gender differentials are reinscribed into novel social arenas and social forms." (Schwemmer et al. 2020).

The use of algorithms to target specific people with political content online, for example, may impact political knowledge, participation, and power differently for men and women. Artificial intelligence systems that recommend digital content based on user behavior can promote false beliefs and enhance political polarization (Harris 2017) (see Chapter 2 on information ecosystems). News dissemination algorithms trained on materials that suggest that women are less interested in or knowledgeable about politics may target women with mommy blogs, workout routines, and fluff pieces, and target men, who have been understood as more likely to give larger amounts to campaigns, with more explicit political content. Studying political targeting and its effects will require some sense of a baseline, for example, understanding what media women and men would have encountered and consumed in the absence of algorithmic targeting. Gender-targeted political advertising may also be contributing to the divergence of public opinion between younger men and women.

The advent of image generation capabilities also creates new opportunities for studying – and new risks for – women online. Artificial intelligence–driven microtargeting techniques could also be used to identify vulnerabilities or specific demographics to target with negative campaigning. Deepfakes, for example, are difficult to differentiate from real videos for the regular consumer and primarily target women (Wagner and Blewer 2019) and those who identify as lesbian, gay, or bisexual (Lenhart et al. 2016).

Deepfake technology has been used notoriously to create pornographic videos that superimpose the face of a target of harassment onto the body of someone in the video. The technology has also been used to disclose private, intimate videos or images without consent – what is often referred to as "revenge porn" (Franks 2017). The use of AI algorithms to recreate images of real people and adapt photos into highly sexualized images (Conger and Dearden 2026) can further fuel harassment of women (Lazard et al. 2025) and other marginalized groups online, as well as alter individual behavior as they become more hesitant to post photos of themselves. Moreover, the use of AI tools to amplify online harassment and misinformation can create a hostile environment that disproportionately drives female politicians to moderate their behavior (Nadim and Fladmoe 2021) and may affect their willingness or ability to maintain political careers. Deepfakes and tools for visual online harassment may reverse the progress in women's representation. Following recent work on the effects of online harassment on gendered behavior (Southern 2026; Wagner 2022), gender scholars should also expand their scope of study to include the effects of LLMs on political behavior, both online and offline.

PROOFS

Even when not used for deepfakes or harassment, AI generation models can act as "amplifiers" of stereotypes. Researchers have found that, in text-to-image generation models, AI generation can reinforce traditional gender roles and reflect existing gender stereotypes about women's role in society. When LLMs were asked to generate images of a person during everyday activities associated with traditional gender roles, such as childcare, shopping, or baking/cooking, they generated images of women. When they were asked to generate images of a person smoking, doing car maintenance, or playing football, they generated images of men (Girrbach et al. 2025). Algorithmic image recognition systems already reproduce bias by focusing much more significantly on the appearance of women over men (Schwemmer et al. 2020), and image generation algorithms likely share the same biases.

This bias extends to images of gendered workplaces (Górska and Jemielniak 2023).

Artificial intelligence models overrepresent men when generating images, even relative to the already existent gender disparity in the health-care space. When asked to produce an image of a physician or doctor, five separate AI models generated images of people who "were more frequently White (82% vs 63%; $P < .001$) and more frequently men (93% vs 62%; $P < .001$) compared with the US physician population." Shockingly, even though women make up 38 percent of physicians in the United States, only 7 percent of the 1,000 images generated were women (Lee et al. 2024). Scholars of mass political discourse and the political behavior of firms should pay attention to the debates around normative AI behavior. Given that many workforces in the United States are increasingly reaching gender parity, *should* AI reflect the current balance of the workforce? Or should AI reflect the workforce as a whole, regardless of the makeup of an individual profession? While these debates are often relegated to ethicists, they are also occurring all across the engineering workforce. The political leanings of the engineers who help to create and train LLMs are reflected in the performance of the models themselves and can magnify social biases given the widespread adoption of AI models in everyday life. Inequality and bias, some scholars have argued, "is not a correctable anomaly but an intrinsic feature of generative AI, rooted in historically uneven data ecologies and design choices" (Kerche et al. 2026). Meanwhile, women across the workforce are slower to adopt AI into their work routines, partially due to their perceptions of the ethics of AI use (Otis et al. 2024). Given the existent gendered data gaps (Criado Perez 2019) and possible political leanings of design choices, AI could come to represent a primary driver of gendered inequality across the labor and political spheres.

(3) **How will AI change jobs and occupational segregation, and how does this impact women's empowerment?**

Artificial intelligence is likely to influence how jobs are characterized and the individuals who will populate those jobs (see Chapter 7 on labor and the economy). Occupational segregation, or the concentration of individuals in the

PROOFS

workforce on the basis of sex, often occurs when new technologies incentivize employers to recast lower earning positions as suitable for women (Reskin and Roos 1990). However, new technologies can also introduce greater opportunities for groups previously excluded from the workforce. The internal combustion engine, for example, contributed to women's participation in the wider political and economic world. Cars gave women the opportunity to leave the home and work, but they also served to reinforce the gendered distribution of labor primarily via their use as another tool for accomplishing house work – for example, for running errands and chauffeuring children (Nichols 2024). Morris (2015) argues that steam power and the age of fossil fuels contributed to a substantial shift in gender roles with the decline of the agrarian economy due to a decline in the "need for muscle power" and a rise in the need for "brainpower and services" women could perform "just as well as men" (102).

For large, inflexible organizations with entrenched occupational hierarchies, AI may face resistance if integration requires retraining a workforce that heavily identifies with their current roles and skill sets. Military organizations, for example, tend to be bureaucratically rigid and often stifle opportunities for change (Rosen 1991, 1–3). This is especially true when emerging technologies introduce roles that fail to align with traditional warfighting (Pindyck 2022). Given the salience of AI to military operations, we might expect military organizations to recast individuals within such roles as masculine "warriors" and potentially rethink systems of reward and ascension to attract sufficient talent.[6]

A pressing issue is when, and in which occupations, autonomous machines will replace human labor (Korinek and Juelfs 2023). In capitalist systems, women's empowerment is linked to their right to enter the labor force, be fairly compensated, and thereby increase their satisfaction with life. Women's empowerment is likewise linked to their class-position in the economy. Artificial intelligence has already threatened routine jobs more than care-oriented jobs. Haircuts and massages are hard to out-source or automate, but factory assembly has proved amenable to machines and robots substituting for human labor. Thus, in the lower classes, women's employment in some parts of the service sector is likely less vulnerable than men's. In white-collar industries, AI adoption originally required skilled employees and, in particular, computer programmers and engineers, a workforce which has skewed male since the 1980s (Abbate 2012).[7] The implementation of AI in hiring and promotion practices can magnify this phenomenon. Artificial intelligence recruitment tools often use a training corpus based on male resumes, favoring "masculine" language and penalizing mentions of the word "woman" (Dastin 2018). Building on the idea of "breadwinner backlash," one hypothesis is that lower class women

[6] Slayton (2021).

[7] Notably, this group is most likely to lose jobs to AI as workforce roles shift from routine coding to the orchestration of system architecture and oversight, requiring engineers to manage, test, and refine AI-generated code.

PROOFS

will suffer in the household as the men of their class face greater layoffs (Clark et al. 2023), while women in the upper classes may suffer in the workforce if they adapt more slowly to the new technology. At the same time, these returns to technical skills may fall over time, setting into motion a decline in male-leaning technical fields.

Notably, recent research suggests that exposure to generative AI can increase public concern over such disruptive effects (Haslberger et al. 2025). From this perspective, increased use of AI at home and at work can also shape public perceptions of how AI technology should be regulated and activate support for social policy measures. The degree to which such support translates into policy outcomes, however, remains to be seen.

6.7 CONCLUSION

Artificial intelligence has transformative potential for political, social, and economic life. In gender and politics, AI not only can be used as a tool for data collection and analysis but also must become a category of analysis itself. With this chapter, we hope to galvanize gender scholars to further study AI's impact on gender and power. The scope of potential biases in AI systems and the technology's revolutionary potential introduce a range of research questions that should interest gender scholars. How can we measure bias in AI systems? How do AI algorithms impact political behavior, the nature of work and occupational segregation of labor, gender relations, and gendered organizational structures? Does AI challenge or reinforce men's advantage in agenda-setting and leadership placement and selection in politics? Will AI "feminize" occupations or will its technological uptake exclude women? How will these dynamics unfold in hierarchical organizations? In politics, is generative AI reinforcing gender role stereotypes in candidate performances? Does selective media exposure polarize men and women? It is too soon to tell whether AI will be a force for gender equality or retrenchment, but research into these and other questions can help guide feminist politics for the next decade.

PROOFS

REFERENCES

Abbate, J. (2012). *Recoding gender: Women's changing participation in computing*. MIT Press.

Aggarwal, N., Matheny, M. E., Shachar, C., Wang, S., & Thadaney-Israni, S. (2022). Artificial intelligence in healthcare. In J. B. Bullock, Y.-C. Chen, J. Himmelreich, V. M. Hudson, A. Korinek, M. M. Young, & B. Zhang (Eds.), *The Oxford handbook of AI governance*. Oxford University Press.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (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* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29* (pp. 4349–4357). Curran Associates. ISBN: 9781510838819.

Boussalis, C., Coan, T. G., Holman, M. R., & Müller, S. (2021). Gender, candidate emotional expression, and voter reactions during televised debates. *American Political Science Review*, *115*(4), 1242–1257. https://doi.org/10.1017/S0003055421000666

Bowdish, L. (2012). American women's struggle to end credit discrimination in the twentieth century. In J. Logemann (Ed.), *The development of consumer credit in global perspective: Business, regulation, and culture* (pp. 103–126). Palgrave Macmillan. https://doi.org/10.1057/9781137062079_6

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, Proceedings of Machine Learning Research (PMLR) (pp. 77–91). Available from https://proceedings.mlr.press/v81/buolamwini18a.html

Cirillo, D., Catuara-Solarz, S., Morey, C., et al. (2020). Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *npj Digital Medicine*, *3*, 81. https://doi.org/10.1038/s41746-020-0288-5

Clark, R., Khoban, R., & Zucker, N. (2023). Breadwinner backlash: The gendered effects of industrial decline [Paper presentation]. American Political Science Association Annual Meeting. Los Angeles, CA.

Cohen, D. K., & Karim, S. (2022). *Does more equality for women mean less war? Rethinking sex and gender inequality and political violence*. Cambridge University Press.

Conger, K., & Dearden, L. (2026, January 9). Grok, Elon Musk's A.I., is generating sexualized images of real people, fueling outrage. *The New York Times*. www.nytimes.com/2026/01/09/technology/grok-deepfakes-ai-x.html

Crawford, K., Dobbe, R., Dryer, T., et al. (2019). *AI Now 2019 report*. AI Now Institute.

Criado Perez, C. (2019). *Invisible women: Data bias in a world designed for men*. Abrams Press.

Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. https://tinyurl.com/3arsn6ea

Dietrich, B. J., Hayes, M., & O'Brien, D. Z. (2019). Pitch perfect: Vocal pitch and the emotional intensity of congressional speech. *American Political Science Review*, *113*(4), 941–962. https://doi.org/10.1017/S0003055419000467

Farrell, H., & Fourcade, M. (2023). The moral economy of high-tech modernism. *Daedalus*, *152*(1), 225–235. https://doi.org/10.1162/daed_a_01982

Franks, M. A. (2017). "Revenge porn" reform: A view from the front lines. *Florida Law Review*, *69*(5), 1251–1337.

Funk, K. D., Paul, H. L., & Philips, A. Q. (2022). Point break: Using machine learning to uncover a critical mass in women's representation. *Political Science Research and Methods*, *10*(2), 372–390. https://doi.org/10.1017/psrm.2021.51

Girrbach, L., Alaniz, S., Smith, G., & Akata, Z. (2025). A large scale analysis of gender biases in text-to-image generative models (arXiv:2503.23398) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2503.23398

PROOFS

Górska, A. M., & Jemielniak, D. (2023). The invisible women: Uncovering gender bias in AI-generated images of professionals. *Feminist Media Studies*, *23*(8), 4370–4375. https://doi.org/10.1080/14680777.2023.2263659

Harris, L. (2017). Overview of artificial intelligence (CRS Report No. IF10608). Congressional Research Service. https://crsreports.congress.gov/product/pdf/IF/IF10608

Haslberger, M., Gingrich, J., & Bhatia, J. (2025). Rage against the machine? Generative AI exposure, subjective risk, and policy preferences. *Journal of European Public Policy*, *32*(3). https://doi.org/10.1080/13501763.2025.2554903

Kerche, F. W., Zook, M., & Graham, M. (2026). The silicon gaze: A typology of biases and inequality in LLMs through the lens of place. *Platforms & Society*, *3*, 1–31. https://doi.org/10.1177/29768624251408919

Korinek, A., & Juelfs, M. (2023). Preparing for the (non-existent?) future of work. In J. B. Bullock, Y.-C. Chen, J. Himmelreich, V. M. Hudson, A. Korinek, M. M. Young, & B. Zhang (Eds.), *The Oxford handbook of AI governance* (pp. 746–776). Oxford University Press.

Lazar, S. (2022). Power and AI: Nature and justification. In J. B. Bullock, Y.-C. Chen, J. Himmelreich, V. M. Hudson, A. Korinek, M. M. Young, & B. Zhang (Eds.), *The Oxford handbook of AI governance* (pp. 198–209). Oxford University Press.

Lazard, L., Capdevila, R., Turley, E. L., Gilfoyle, K., & Stavropoulou, N. (2025). Deepfake technology and gender-based violence: A scoping review. *Trauma, Violence, & Abuse*, 1–16. https://doi.org/10.1177/15248380251384271

Lee, S. W., Parikh, R. B., Sapkota, A., et al. (2024). Demographic representation of generative artificial intelligence images of physicians. *JAMA Network Open*, *7*(8), e2425993. https://doi.org/10.1001/jamanetworkopen.2024.25993

Lenhart, A., Ybarra, M., & Price-Feeney, M. (2016). *Nonconsensual image sharing: One in 25 Americans has been a victim of "revenge porn."* Data & Society Research Institute. https://tinyurl.com/neujfej4

Littman, M. L., Ajunwa, I., Berger, G., et al. (2021). Gathering strength, gathering storms: The one hundred year study on artificial intelligence (AI100) 2021 study panel report. Stanford University. https://tinyurl.com/yd98xnmv

Meguid, B. M., Coffé, H., Weeks, A. C., & Kittilson, M. C. (2025). Strategic inclusion without transformation: How populist radical right parties engage with women's interests. *Comparative Political Studies*. Advance online publication. https://doi.org/10.1177/00104140251381760

Miller, M. G., & Sutherland, J. L. (2023). The effect of gender on interruptions at congressional hearings. *American Political Science Review*, *117*(1), 103–121. https://doi.org/10.1017/S0003055422000260

Morris, I. (2015). *Foragers, farmers, and fossil fuels: How human values evolve*. Princeton University Press.

Mundy, L. (2017, April). Why is Silicon Valley so awful to women? *The Atlantic*. www.theatlantic.com/magazine/archive/2017/04/why-is-silicon-valley-so-awful-to-women/517788/

Nadim, M., & Fladmoe, A. (2021). Silencing women? Gender and online harassment. *Social Science Computer Review*, *39*(2), 245–258. https://doi.org/10.1177/0894439319865518

PROOFS

Ngan, M., & Grother, P. (2015). Face recognition vendor test (FRVT): Performance of automated gender classification algorithms (NIST Interagency Report 8052). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8052

Nichols, N. A. (2024). *Women behind the wheel: An unexpected and personal history of the car*. Simon and Schuster.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Otis, N. G., Delecourt, S., Cranney, K., & Koning, R. (2024). Global evidence on gender gaps and generative AI (Working Paper No. 25-023). Harvard Business School. https://doi.org/10.31219/osf.io/h6a7c

Pindyck, S. (2022). *Are gender inclusive militaries better at integrating disruptive technologies?* University of California Institute on Global Conflict and Cooperation.

Reskin, B. F., Roos, P. A., Donato, K. M., et al. (1990). *Job queues, gender queues: Explaining women's inroads into male occupations*. Temple University Press.

Rickman, S. (2025). Evaluating gender bias in large language models in long-term care. *BMC Medical Informatics and Decision Making*, *25*, 274. https://doi.org/10.1186/s12911-025-03118-0

Ridgeway, C. L. (2011). *Framed by gender: How gender inequality persists in the modern world*. Oxford University Press.

Rosen, S. P. (1991). *Winning the next war: Innovation and the modern military*. Cornell University Press.

Schwemmer, C., Knight, C., Bello-Pardo, E. D., et al. (2020). Diagnosing gender bias in image recognition systems. *Socius: Sociological Research for a Dynamic World*, *6*, 1–13. https://doi.org/10.1177/2378023120967171

Sjoberg, L., Kadera, K. M., & Thies, C. G. (2018). Reevaluating gender and IR scholarship: Moving beyond Reiter's dichotomies toward effective synergies. *Journal of Conflict Resolution*, *62*(4), 848–870. https://doi.org/10.1177/0022002716669207

Slayton, R. (2021). What is a cyber warrior? The emergence of U.S. military cyber expertise, 1967–2018. *Texas National Security Review*, *4*(1), 61–96.

Smith, G., & Rustagi, I. (2021). When good algorithms go sexist: Why and how to advance AI gender equity. *Stanford Social Innovation Review*. https://doi.org/10.48558/a179-b138

Sorokovikova, A., Chizhov, P., Eremenko, I., & Yamshchikov, I. P. (2025). Surface fairness, deep bias: A comparative study of bias in language models. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)* (pp. 206–227). Association for Computational Linguistics.

Southern, R. (2026). Does fear of harassment reduce women's online political participation? Evidence from the British Election Study across three UK general elections. *European Journal of Politics and Gender*, *9*(1), 223–243. Retrieved July 10, 2026, from https://doi.org/10.1332/25151088Y2024D000000027

Spinks, C. N. (2020). Contemporary housing discrimination: Facebook, targeted advertising, and the fair housing act. *Houston Law Review*, *57*(4), 925–970.

Tang, C., Li, S. K., Hu, S., Zeng, F., & Du, Q. (2025). Gender disparities in the impact of generative artificial intelligence: Evidence from academia. *PNAS Nexus*, *4*(2), page 591. https://doi.org/10.1093/pnasnexus/pgae591

PROOFS

Teele, D. L., Kalla, J., & Rosenbluth, F. M. (2018). The ties that double bind: Social roles and women's underrepresentation in politics. *American Political Science Review*, *112*(3), 525–541. https://doi.org/10.1017/S0003055418000217

Vredenburgh, K. (2022). Fairness. In J. B. Bullock, Y.-C. Chen, J. Himmelreich, V. M. Hudson, A. Korinek, M. M. Young, & B. Zhang (Eds.), *The Oxford handbook of AI governance* (pp. 129–148). Oxford University Press.

Wagner, A. (2022). Tolerating the trolls? Gendered perceptions of online harassment of politicians in Canada. *Feminist Media Studies*, *22*(1), 32–47. https://doi.org/10.1080/14680777.2020.1749691

Wagner, T. L., & Blewer, A. (2019). "The word real is no longer real": Deepfakes, gender, and the challenges of AI-altered video. *Open Information Science*, *3*(1), 32–46. https://doi.org/10.1515/opis-2019-0003

Whose genomics? (2019). *Nature Human Behaviour*, *3*(5), 409–410. https://doi.org/10.1038/s41562-019-0619-1

PROOFS

7

AI's Economy and Its Political and Institutional Consequences

Carles Boix, Michael Becher, Valentina González-Rostani, and Daniel Stegmueller

PROOFS

7.1 INTRODUCTION

Artificial intelligence (AI) has been increasingly hailed as the ultimate technological frontier – a tool or set of tools capable of yielding explosive productivity gains and replacing all human labor in the next few years or decades. At the same time, however, it has triggered widely divergent assessments about its social and political consequences. While some celebrate it as a liberating instrument that will usher in a 24/7 leisure society, others predict that it will lead to a dramatically unequal society, with a split between a few ultra-rich and an impoverished multitude, and to a powerful state capable of controlling all the minutiae of its citizens' daily lives.

In this chapter, which examines the political and institutional consequences of the economic changes brought about by AI, we arrive at a more conservative, middle-of-the-road position. This is probably the case for two reasons. First, we rely on evidence gathered during the short period of time that has elapsed since the development and actual implementation of AI. Second, we find substantial variation in researchers' assessments of its impact. Nevertheless, going beyond our description (based on current research) of AI's short-term effects, we also provide a conceptual discussion of its (likely stronger) political and social repercussions over the medium and long run.

The chapter can be summarized around seven broad claims:

First, whereas the information technology revolution benefited highly skilled workers and polarized labor markets, AI may replace nonmanual, nonroutine tasks preferentially done by highly educated individuals. This prediction is, however, uncertain for at least two reasons. Current (and widely used) measures and indexes developed to estimate the exposure

Prepared for the APSA Task Force on AI and Political Science.

Tucker_9781047774734_C007.indd 209 15-09-2026 07:03:21

and vulnerability of labor tasks to AI are weakly correlated with each other, yielding rather contradictory forecasts about which occupations are more likely to be replaced in the future. In addition, the methods and procedures of AI, which researchers normally use to assess its progress, are different from its actual application to management and production processes.

Second, the effect of AI on employment, collective bargaining, and wages is likely to be conditional on the structure of labor institutions.

Third, public attitudes toward AI are in a state of flux. Even though a significant fraction of the public worries about the labor displacement effects of AI, a majority in advanced countries still view AI as a positive development. Reactions to AI, which are more negative among young cohorts, at least on labor-related issues, are strongly susceptible to the type of questions and treatments developed in the surveys and, in at least one study, highly affected by some personality traits of respondents.

Fourth, although most economic researchers and survey analysts have focused on the impact of AI on labor markets, AI's economic (and, therefore, political) effects seem to be as – or even more – consequential on capital. The use and training of foundation models and the computational demands that come with those processes have accelerated capital concentration and reduced capital mobility in the technological sector. Those effects have increased, in turn, the lobbying power of technological firms as well as their incentives to capture the regulator.

Fifth, while information technologies are associated with the political polarization of the last decades, we cannot discard the possibility that, if AI ends up replacing nonroutine jobs, it could foster a process of political de-polarization. Yet, perhaps paradoxically, this effect may not stabilize electoral politics due to the negative impact of AI on the relative position of individuals and on younger cohorts entering the labor market.

Sixth, AI-driven economic changes may have negative consequences for democracy. The combination of labor substitution (and a concomitant reduction of wages) and growing capital concentration (particularly in the technological sector) may jeopardize the social consensus needed to accept democratic elections. Still, higher productivity and economic growth, if distributed fairly, could reduce social and political conflict, even to the point of neutralizing any potential de-democratizing effects of AI.

Finally, AI may affect not only the within-country distribution of economic gains and losses (triggering a corresponding electoral realignment). It could reverse recent economic trends toward cross-country economic convergence (related to globalization) and widen the technological gap between developed and developing economies.

The overall structure of the chapter is as follows. After presenting existing evidence on the changes triggered by AI on capital and labor markets, we outline its political consequences. Section 7.2 ("Labor Market") discusses

PROOFS

three topics related to AI and labor: current evidence on the labor-augmenting and labor-substituting effects of AI; the nature and robustness of AI exposure indexes; and the potential mediating effects of labor unions and wage bargaining institutions on AI's economic impact. Section 7.3 ("AI and Public Opinion") explores the public's perceptions toward AI. Section 7.4 ("Capital Concentration and Market Power") describes the impact of AI on the nature and concentration of capital. Section 7.5 ("AI and Democracy") considers how an AI-induced transformation of the labor market and capital may shape democratic politics – both triggering an electoral realignment and destabilizing political institutions in general. Most of the research and discussion on AI's consequences has been done so far for advanced economies. However, the spread of AI should have important effects on the Global South, a topic we discuss in Section 7.6 ("Economic and Political Backsliding Across the World?"). The concluding section suggests future avenues of research for the social science community and offers some policy recommendations.

7.2 LABOR MARKET

7.2.1 Substitution and Augmentation

The task model of production offers a convenient way to study how automation and AI affect work. According to this view, production is organized as a collection of routine and nonroutine tasks that can be performed by capital or labor, with labor hired either domestically or offshore.[1] Firms assign each task to the cheapest feasible input, given technology and relative prices.

Over the last few decades, several waves of digitalization and computerization substituted machine-based procedures for routine tasks while raising the value of nonroutine problem-solving and interpersonal work.[2] The structure of the labor market shifted accordingly. While the need for middle-skill clerical and production tasks fell, the demand for high-skill analytic activities and low-skill service work went up (Autor et al., 1998, 2002, 2003). Job polarization followed. Employment shrank in middle-wage occupations while jobs grew at the top and bottom of the wage distribution (Goos & Manning, 2007; Goos et al., 2009; Autor, 2013, Humlum, 2019; Jaimovich & Siu, 2019; Hoffmann et al., 2020), causing a shift in the US wage-structure (Acemoglu & Restrepo, 2022). Regions and workers more exposed to robots or other forms of automation, especially in manufacturing, experienced job losses and weaker wage growth. That resulted in broader local spillovers in the form of population decline and fewer high-skill job opportunities (Chiacchio et al., 2018; Graetz & Michaels, 2018; Acemoglu & Restrepo, 2020; Sarto et al., 2025). Individual-level studies

[1] See Acemoglu and Autor (2011) and Autor (2013) and the review in Acemoglu and Restrepo (2019).

[2] Digitalization refers to the adoption of ICT (Kurer & Gallego, 2019) and computerization to job automation by computer-controlled equipment (Frey & Osborne, 2017, p. 254).

PROOFS

show higher separation risk and persistent earnings losses after automation shocks, with many displaced workers either accepting lower paying jobs that still use their skills or leaving the labor force, particularly among older workers (Bessen et al., 2025; Lerch, 2021; Braxton & Taska, 2023).[3]

Technological change arguably contributed, too, to a shift in the capital and labor shares of national income. By 2022, the labor share of US national income was at its lowest since 1929. After peaking in 1946, it fell by 0.2 percentage points per decade until 1970. Coinciding with the entry of information and digital technologies, its decline accelerated to 1.6 percentage points loss per decade from 1971 to 1995 and by 1.8 percentage points per decade after 1996. Labor shares fell across almost all industries since 1987. In the information industry, the labor share for 1987–2021 averaged 43 percent – the fourth lowest among twenty industries in the economy and only higher than real estate, utilities, and mining and oil (Karabarbounis, 2024). By contrast, in 2024 after-tax corporate profits were equivalent to 11.5 percent of GDP, almost twice their level in the 1980s.[4]

New developments in AI have the potential to extend automation from routine activities into complex cognitive and interactive tasks and, potentially, into full workflows with limited human oversight (Susskind & Susskind, 2022; Autor, 2024). These possibilities have sparked a wide debate about AI's long-run labor-market effects.

For some, AI will eventually substitute human labor, resulting in a steep decline in labor's income share, which, at some point, will approach zero (Restrepo, 2025). For others, AI will mainly augment human labor, raise workers' productivity, or enable them to perform more complex responsibilities. Here, AI is even foreseen as a tool to rebuild parts of the middle class by extending expert capabilities to more workers (Autor, 2024). For example, nurse practitioners – already performing tasks once limited to physicians thanks to digital records and improved communication systems – could see their clinical judgment strengthened as AI supports an expanding scope of practice. A third, more sanguine position, sees AI as a standard technology that reallocates tasks and shifts relative demands without fully displacing workers (Narayanan & Kapoor, 2025).

Either way, the substitution and augmentation effects of AI will likely be conditional on the pace of technological progress and on the specificity and availability of the required complementary skills. Rapid technological advances – either exceeding the capacity of education and training systems to adapt to them or rewarding very narrow technical skills – will tend to push

[3] The net effect of pre-AI automation on employment seemed to depend on how quickly new tasks and sectors expand and on the extent of reallocation within and across firms (Autor & Salomons, 2018; Acemoglu & Restrepo, 2019; Dauth et al., 2021; Koch et al., 2021).

[4] Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/W273RE1A156NBEA. Accessed on February 14, 2026.

PROOFS

production toward labor substitution and downward wage pressure. Slower or more predictable change, including skill demands that can be met through incremental retraining, will create more room for workers to adjust and for AI to act as a productivity aid, rather than a replacement tool. In addition, governance and workplace norms may slow or reshape the adoption of AI advances in high-stakes settings, even when the technical capability exists.

Current empirical work on the labor market consequences of AI takes two forms: studies measuring the potential exposure or vulnerability of current occupations to AI as a function of the latter's characteristics; and empirical research on realized outcomes.

7.2.1.1 Measuring Exposure to AI

Most recent measures tying AI capabilities to job requirements – whether through patents, benchmark tasks, or model-task mappings – tend to find higher levels of AI exposure among higher education, higher wage occupations (Webb, 2020; Felten et al., 2021). Using large-language-model task mappings, Manning et al. (2023) estimate that roughly 80 percent of workers have at least 10 percent of their tasks exposed, and that between one-fifth and one-third may face disruption to half or more of their tasks (Kinder et al., 2024). Macro projections similarly suggest that up to 8 percent of total hours could be automatable by generative AI by 2030 (Ellingrud et al., 2023).

The estimated exposure or vulnerability to AI seems to be uneven across skills, industries, regions, and age groups (Webb, 2020; Felten et al., 2021; Brynjolfsson et al., 2025; Huang, 2025) and, crucially, very dependent on how exposure is defined – whether at the occupation or task level, whether based on technical capability or observed deployment, and whether augmentation is counted as exposure.[5]

Figure 7.1 compares three popular measures of AI vulnerability (Brynjolfsson et al., 2018; Webb, 2020; Felten et al., 2023). We matched each measure to the Standard Occupational Classification (SOC) 2010 classification yielding triplets of AI-exposure measures for 775 harmonized occupations in the US. The main diagonal of Figure 7.1 shows the distribution of each measure, while the lower off-diagonal cells provide bivariate scatterplots. Their distribution differs and the scatterplots do not reveal clear patterns of covariation. This is confirmed by the low-rank correlations (Kendall's τ) as well as low-distance correlations (which also capture patterns of nonlinear dependence; Székely et al., 2007). Both measures are reported in the upper off-diagonal cells. Table 7.A illustrates a larger number of widely used indicators since the early 2000s and shows how different measurement choices yield sharply different

[5] Importantly, however, these exposure measures represent technical upper bounds. They indicate where AI could substitute for or augment tasks if firms re-engineer workflows not how much substitution is occurring now or how fast it will unfold (Brynjolfsson et al., 2021; McElheran et al., 2025).

PROOFS

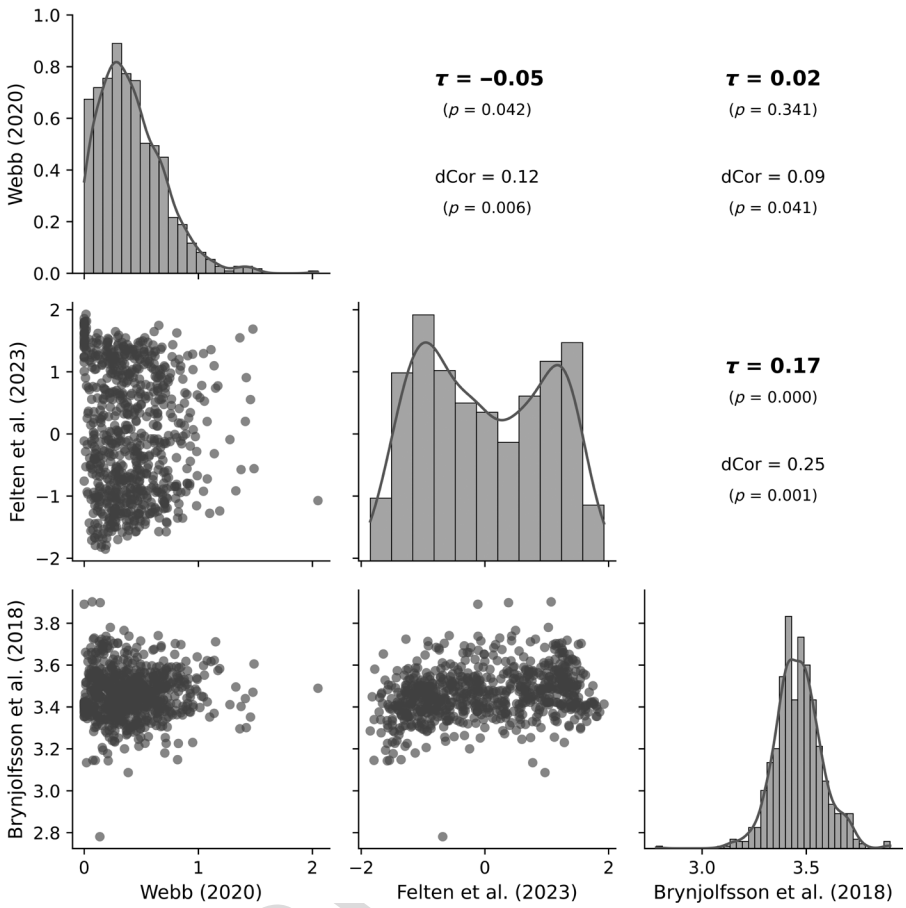


FIGURE 7.1 Comparison of three popular AI-exposure measures.

headline numbers (Arntz et al., 2017; Frey & Osborne, 2017; Webb, 2020; Felten et al., 2021; Manning et al., 2023).

We interpret these differences as a sign that each vulnerability index is capturing different aspects of the same underlying phenomenon (and measuring the latter by emphasizing different dimensions related to work). If this is the case, the predicted economic and political implications of these measures are heterogeneous and, as a matter of fact, substantially uncertain. We consider the economic effects here. We turn to the political consequences in Section 7.5.1 ("The Electoral Arena").[6]

[6] This variability in AI exposure indexes has implications for empirical practice, too. If they indeed capture different aspects of the same phenomenon (as opposed to being noisy measures of an underlying latent variable), empirical analyses should use them in combination and not as

Tucker_9781047774734_C007.indd 214 15-09-2026 07:03:23

TABLE 7.1 *Sector-specific exposure measures*

	Felten		Brynjolfsson		Webb	
Sector	**Value**	**Rank**	**Value**	**Rank**	**Value**	**Rank**
Finance and insurance	0.861	1	3.540	1	0.376	7
Educational services	0.839	2	3.488	6	0.310	10
Information	0.634	3	3.530	3	0.429	4
Public administration	0.435	4	3.487	7	0.344	9
Professional, Scientific, and Technical Services	0.344	5	3.498	5	0.425	5
Wholesale trade	0.094	6	3.513	4	0.374	8
Retail trade	0.012	7	3.539	2	0.296	12
Other services (except Public Administration)	−0.138	8	3.460	9	0.301	11
Manufacturing	−0.372	9	3.477	8	0.471	2
Arts, entertainment, and recreation	−0.407	10	3.454	10	0.179	13
Construction	−0.743	11	3.439	12	0.438	3
Transportation and warehousing	−0.799	12	3.450	11	0.405	6
Agriculture, forestry, fishing	−1.117	13	3.414	13	0.636	1

To get a sense of the economy-wide implications of AI exposure, Table 7.1 shows (private-sector) employment-weighted averages of occupational exposure to AI (and their rank order) for thirteen major sectors of the economy. The values are calculated by matching different AI exposure measures to national employment estimates by occupation and the North American Industry Classification System (NAICS) two-digit sector from the Bureau of Labor Statistics' Occupational Employment and Wage Statistics Survey[7] and then aggregating them to thirteen major sectors.

While the indexes of both Brynjolfsson et al. and Felten et al. rank finance and insurance as highly likely exposed, Webb's AI measure ranks them only seventh and instead implies that agriculture and manufacturing are highly vulnerable. On the other hand, all three measures agree that the information sector and professional, scientific, and technical services are relatively highly exposed. It is also noteworthy that, according to the measure by Felten et al. (2021) and especially Brynjolfsson et al. (2018), retail and wholesale trade face sizable exposure, indicating that AI as a general-purpose technology (Bresnahan & Trajtenberg, 1995; Brynjolfsson et al., 2021) is likely to penetrate widely throughout all economic sectors.

"robustness tests" of each other. For a study that distinguishes between GenAI exposure (or the theoretical technical applicability of GenAI) and AI-substitution vulnerability (or the practical likelihood that employers replace workers with AI) and considers their joint effect on labor replacement, see Liu et al. (2025).

[7] See U.S. Bureau of Labor Statistics, www.bls.gov/oes. We use May 2023 estimates.

PROOFS

As Narayanan and Kapoor (2025) emphasize, however, current indicators should be treated as guides rather than predictions, since they often conflate progress in methods with progress in applications and diffusion. Frontier benchmarks showing that models can pass professional exams or solve self-contained coding tasks do not reveal whether firms will trust these systems with high-stakes legal work, complex software development, or safety-critical operations. Likewise, existing uplift studies mainly document incremental productivity improvements, rather than large-scale automation of occupations. What remains uncertain, therefore, is not only how quickly new tasks will emerge, how broadly augmentation will diffuse, and how far institutions and policies will go in smoothing worker transitions, but also which empirical indicators best capture these processes and how far ahead they can be projected without overstating the scale of disruption.

7.2.1.2 Realized Effects of AI

Compared to digital technologies and computer-controlled equipment, AI has been deployed for a limited number of years. This logically affects both the number and robustness of any studies on its labor market effects. Overall, however, the evidence so far points to some productivity gains, substitution in routine and possibly many nonroutine tasks, and rising demand for AI-complementary skills.

On the one hand, recent work finds AI to be restructuring employment patterns in two directions. First, areas more exposed to AI-related advances have seen weaker employment and wage growth among workers whose tasks are close substitutes for current systems (Huang, 2025). Still, this net displacement has been estimated to be smaller once new tasks and complementarities are considered, particularly in nonroutine jobs (Kogan et al., 2023; Pizzinelli et al., 2023). Brynjolfsson et al. (2025) find that AI's replacement impact may be operating mostly among young cohorts, arguably because they have not been able to develop yet the kinds of tacit knowledge and contacts gained while working that may make up for AI-led substitution. Examining 285 million job postings from 2018 to 2025, Liu et al. (2025) find that occupations vulnerable to artificial intelligence experienced a fall in demand following the release of ChatGPT in the fall of 2022. By contrast, a recent analysis of job posting data in Germany, the Netherlands, the United Kingdom and the United States shows no evidence that AI is behind a fall in early career hirings, instead attributing the latter to rising interest rates in the middle of 2022 (Burn-Murdoch & O'Connor, 2026a).

Second, current surveys indicate that AI-related skills are increasingly valued over prior job experience in many industries (Microsoft, 2024). Skill profiles with AI-related capabilities have enjoyed a wage premium of roughly 5–10 percent in many countries (Alekseeva et al., 2021) and even higher levels in some emerging markets (Copestake et al., 2023). Firm-level studies show that adopters reorganize toward more educated, STEM-intensive workforces, with

PROOFS

fewer mid-level managers, higher revenue, and higher valuations (Babina et al., 2023, 2024). However, most early adopters do not immediately reduce headcount. An Organization for Economic Cooperation and Development (OECD) survey reports that only about one-quarter of AI-using firms cut employment, with many redeploying workers to new tasks (Milanez, 2023).

On the other hand, AI also appears to augment human labor, although these effects vary widely with worker experience and type of task. Studies of human-AI teams in customer support show that access to a generative AI–assistant reduced resolution times by about 14 percent, with the largest gains for less experienced workers (Brynjolfsson et al., 2025). Experimental settings for writing tasks find higher-quality drafts produced more quickly and a narrowing of performance gaps between lower- and higher-skill workers (Noy & Zhang, 2023). In software development, programmers using tools such as GitHub Copilot complete coding tasks substantially faster than without them (Peng et al., 2023). However, there is also randomized controlled trial evidence suggesting slowdown effects for experienced developers (Becker et al., 2025).[8] In settings with highly trained specialists, such as expert radiologists, AI decision-support has so far had limited effects on accuracy because professionals choose not to rely on its recommendations (Agarwal et al., 2023; Yu et al., 2024).

We conclude this subsection by making two additional observations. First, we still know little about how AI-driven employment decisions interact with bargaining institutions, competition policy, and data governance. This makes it difficult to read long-run wage and employment trajectories from the behavior of early adopters alone (Narayanan & Kapoor, 2025).

Second, most of the discussion so far concerns software and digital services. As AI becomes embedded in physical equipment and robotics, "embodied AI" may result in the automation of manual and service work, with fundamental consequences in terms of substitution and augmentation across cognitive and physical tasks and, as we discuss in Section 7.5, for politics.

7.2.2 Labor Unions and Worker Mobilization

Besides shaping labor demand, technological change (and AI should be no exception here) may impact labor market institutions, that is, the strength of labor unions (trade unions in British English) as well as the level and reach of collective wage bargaining. This may affect, in turn, wages and working conditions (Farber et al., 2021; Jäger et al., 2025), inequality (Iversen, 1999; Hall & Soskice, 2001), and democratic politics (Leighley & Nagler, 2007; Ahlquist, 2017; Kerrissey & Schofer, 2013; Rosenfeld, 2019; Becher & Stegmueller, 2021; Hertel-Fernandez, 2025; Kaplan & Naidu, 2025; Yan, 2025).

[8] Burn-Murdoch and O'Connor (2026b) report an upward inflection point in software productivity by the end of 2025, coinciding with the launch of ClaudeCode and OpenAI's Codex.

PROOFS

Because it is too early to give anything close to a conclusive empirical answer about the consequences of an increased use of AI at work on workers' interests and in their ability to organize in unions, in the remainder of this subsection we discuss several models emphasizing the negative and positive effects that AI may have on wage bargaining and union labor membership sequentially. We conclude by pointing to research on the regulation of AI in collective bargaining agreements.

7.2.2.1 AI May Weaken Labor Unions

Artificial intelligence may weaken labor unions through two channels: skill-biased technological change and labor substitution in imperfectly competitive labor markets.

Artificial intelligence's impact on labor market institutions likely depends on the type of skills and occupations that are more vulnerable to these new technologies. If AI rewards the more educated, existing theories of skill-biased technological change and unions' performance, developed in the 1980s and 1990s (e.g., Pontusson & Swenson, 1996; Iversen, 1999), will continue to apply in the age of AI. According to that view, a shift from Fordist to diversified quality production undermined centralized wage bargaining and solidaristic wage policy pursued by unions. By widening the productivity gap between workers in different sectors or firms, skilled workers who benefited from technological change found it harder to stomach the costs of wage compression, which benefited lower skilled workers. The distributive effects of technological change increased the incentives of technology winners to push for a shift from centralized (nationwide) to industry or even company-level wage bargaining between firms and unions.[9]

Standard task models, which capture the idea that automation may result in labor substitution, assume that all markets are perfectly competitive. Nonetheless, this unrealistic assumption, which brackets the possibility of collective bargaining over wages and working conditions, has been challenged in recent task-based models analyzing the impact of automation on labor unions. Based on a recent extension of the task framework that incorporates labor market search frictions, Leduc and Liu (2024) argue that automation can reduce union power during business cycle downturns. Examining how exogenously given wage rents shape firms' decision to automate, Acemoglu and Restrepo (2024) argue that automation will target high rent tasks. If unions are indeed concentrated in such tasks or more likely to target them for mobilization, firms may have incentives to use AI to pursue inefficient automation in part to weaken existing unions or keep them out of the company's labor force to begin with. Relatedly, Golden (1997) argues that firms that downsize their

[9] The breakdown of Sweden's centralized wage setting system at the time has been explained with reference to this framework (Pontusson & Swenson, 1996). For a formalization of this logic in the private sector, see Acemoglu et al. (2001).

PROOFS

workforce due to technological change or business cycles sometimes want to target dismissals to union members and their workplace leaders. Moreover, employers may use AI to increase the segmentation of the workplace in global value chains, through outsourcing, subcontracting, temporary work, and less worker autonomy (Doellgast et al., 2025).

7.2.2.2 AI May Strengthen Labor Unions

Contrary to our previous discussion, the skill-mediated impact of AI may strengthen labor unions and collective wage bargaining. If it is highly educated individuals who turn out to be more exposed to AI (in line with the Felten and Webb indexes plotted in Figure 7.4 later in the chapter) than less educated workers, support for union-led wage bargaining may increase and converge across all workers.

In addition to the mediating effect of AI on skills, a number of models examine the positive impact of AI on labor unions due to the specific nature of the capital investments AI entails. Building on the task framework and considering an open economy setting with imperfect competition, Becher and Stegmueller (2025) suggest that automation can strengthen (as well as weaken) worker mobilization and bargaining power. In a global economy, the threat to relocate production often gives the upper hand in collective bargaining to firms. However, investments in automation, such as AI-enabled robots, process automation, or data centers powering AI applications, can, under certain conditions, reduce firms' relocating threat. If the productivity gains from automation either depend on co-specific skills in nonautomated tasks or rest on complementary services such as robot integrators, and/or if the cost of relocating and running the machines abroad are relatively high, workers' power at the bargaining table goes up. That increases, in turn, the gains from mobilizing workers. By contrast, if those conditions are not met, and AI replaces unionized labor, AI-driven automation will weaken unions.

Quantitative evidence on the impact of AI on unionization is still scarce. Prior research on twenty-one OECD countries from 1970 to 2019 finds that routine-biased technological change is associated with lower union density (Meyer, 2019). Using an instrumental variable strategy based on robot exposure in the same industries in other advanced economies, recent research finds that robotization reduces union density in Europe (Agnolin et al., 2025) and the US (Balcázar, 2023). After confirming that the overall effects of AI on unionization are negative, Becher and Stegmueller (2025) find significant heterogeneity of effects that is not explained by observable characteristics.

While AI is increasingly used in robotics and process automation, the data on robotization exposure used in these studies is not restricted to AI, and findings on robots may not travel to AI used in other areas, such as algorithmic management or services. Combining survey data on Western Europe (2012–24) with an occupational measure of exposure to new digital technologies across six different categories, Agnolin and González-Rostani (2025) find

PROOFS

that the level of unionization is correlated with the type of technology. While platform-based tools (e.g., for food delivery) are associated with lower unionization, others, such as exposure to machine learning or embedded systems, are associated with higher unionization.

7.2.2.3 Negotiating the Deployment of AI

Qualitative and quantitative work has started to explore how unions negotiate the use of AI at work, through collective bargaining with employers as well as broader social dialogue.

In what may be the most comprehensive case study to date, Doellgast et al. (2025) analyze efforts around the globe to encourage the use of AI to complement rather than substitute labor, to negotiate algorithmic management in a way that protects workers' rights and autonomy, and limit AI-enabled fragmentation of the workforce. It underscores that, besides the question of whether AI displaces or augments labor, another core issue is how AI is deployed for monitoring workers and making managerial decisions related to hiring, scheduling, or pay. In this respect, AI can be used mainly as a tool to discipline workers or to increase their autonomy. The study argues that constraints on employer exit, institutional support for worker voice, and strategies of inclusive labor solidarity foster bargains that encourage a more worker-friendly approach to AI adoption (also see De Stefano & Doellgast, 2023).[10]

7.3 AI AND PUBLIC OPINION

How does the perceived or actual impact of AI on the workplace shape people's policy preferences and political behavior? For more than a decade, political economists and other social scientists have examined how technological shocks to labor markets based on information technology, computerization, and, increasingly, AI are shaping (or not) democratic mass politics (for an excellent review, see Gallego & Kurer, 2022).

As with earlier technological changes, we should expect AI to result in winners and losers. While not meant to be an exhaustive review, this section highlights several directions of a still incipient research on AI-driven economic vulnerabilities and public opinion. Here as well as in the discussion in Section 7.5, it is important to keep in mind that current research in economics still is uncertain about the labor market risks and distributive effects of AI.

Several years into the AI boom, a majority of survey respondents in advanced economies say that recent digital technologies, including AI, have

[10] In the US, unions in the entertainment sector have negotiated collective bargaining agreements for screenwriters and actors that regulate specific uses of generative AI for filmmaking and video-game production. Agreements were only reached after months-long strikes. More broadly, issues of data privacy and worker autonomy are being contested and negotiated across countries and sectors.

PROOF

a positive impact on the economy. However, many also declare that robots and AI will displace more jobs than they create. And a sizable minority is concerned that their job may be displaced (e.g., Busemeyer et al., 2023; European Commission, 2024; Green et al., 2025). A recent survey asking respondents to evaluate AI-driven automation in 940 occupations shows that the American public supports automating 30 percent of occupations. That proportion doubles to 58 percent of occupations when AI is described as outperforming humans at lower cost (Friis and Riley, 2025).

Consistent with some evidence of AI's impact on the labor market (reviewed in Section 7.2), some studies show that the perceived labor market vulnerability to AI is higher among young people. For example, according to a Harvard Youth Poll from 2025, 44 percent of Americans aged eighteen to twenty-nine believe that AI will take away economic opportunities, 22 percent are unsure, less than 20 percent expect gains, and the rest expect no change (IOP, 2025). Surveys across the board also find AI anxiety among highly educated professionals in occupations with no or little exposure to previous waves of technological change.

Given that the adoption of AI at work is uneven, incomplete, and of course likely correlated with both observed and unobserved characteristics of workers, firms, location and so on, randomized experiments constitute an active strand of research. Overall, they capture that public attitudes regarding AI and policies designed to address labor market effects seem malleable to information, elite cues, and experience.

Researchers have explored different treatments and outcome variables. At the risk of oversimplifying, one can distinguish at least three approaches. One common experimental approach randomizes information about AI's potential labor market impacts. Zhang (2022) reports that, in a United States sample, information treatments about the general or occupation-specific AI risks raise the perceived threat of job displacement but have no effects on policy preferences regarding redistribution, immigration, or trade. However, other researchers find that providing information about AI risk increases support for social insurance and redistribution, at least if paired with rhetoric critical of tech elites (Jeffrey, 2021), for regulation, with effects varying by information source (Mitts & Ravis, 2025) or AI knowledge (Heinrich & Witko, 2025), and for protectionist preferences such as trade restrictions (Gonzalez-Rostani, 2024). In an intensive information design, workers across ninety-eight occupations exposed to AI receive an occupation-specific video treatment about the capabilities of generative AI to perform key tasks in their occupation (Anelli et al., 2025). Fielded in Germany, Italy, and the US, the study finds that treatment effects on policy preferences and political attitudes are heterogeneous and vary depending on whether respondents feel threatened by or optimistic about AI.

A second type of experiment explores whether the experience of using AI at work shapes people's policy preferences. For example, Haslberger et al. (2025) randomize whether participants can use generative AI or not to

PROOFS

complete work-related writing tasks in a UK sample of working-age adults, finding that experiential exposure to AI does not increase the perceived risk of job displacement and that it increases the belief that AI will be beneficial to them and society more broadly. Turning to policy preferences, the treatment reduces support for regulating AI but increases support for developing social measures to counter the effects of AI. In an earlier study focused on a sample of college-educated professionals in AI-exposed occupations, Noy & Zhang (2023) find increased concern about job displacement as well as heightened AI-optimism. Also employing an experiential treatment but focusing on management decisions rather than automation risk, related research randomizes whether platform workers interact with an algorithmic or a human boss (Margalit & Raviv, 2025). It finds that workers' personal experience with algorithmic decision-making affects workers' task satisfaction and effort, though not their policy preferences toward AI-based algorithms.

Finally, several experiments expose survey respondents to (hypothetical) scenarios about a firm's decisions to automate jobs and thus lay off workers. Studies focused on the manufacturing sector in different countries find that people are more sensitive to jobs lost to offshoring than to automation (Borwein et al., 2024; Kuo et al., 2024). Subsequent work distinguishes between whether the technology is domestic or foreign (Chaudoin & Mangini, 2025) and conducts conjoint experiments where subjects are either assigned to an AI or offshoring treatment and then complete a small number of rating tasks involving attributes such as positive and negative consumer prices and job gains or losses, intended to capture the heterogeneous effects of AI adoption (Magistro et al., 2025). Respondents are found to be as sensitive or more sensitive to price changes than employment shifts, more responsive to job losses than job gains, and to generally favor AI over offshoring. This relationship holds across the partisan divide in the United States. While Democrats are more supportive of AI, in general, the reaction to *changes* (e.g., job losses) is similar for Republicans and Democrats.

PROOFS

7.4 CAPITAL CONCENTRATION AND MARKET POWER

While discussing the substitution effect of AI technologies, Section 7.2.1 ("Substitution and Augmentation") already pointed to their potential impact on the income share of capital and labor. In this section, we turn to examine its repercussions on both the concentration of capital (in the AI industry as well as the economy in general) and its nature.

7.4.1 AI's Technological Requirements

Powered by the algorithmic revolution and cloud computing, the IT industry coalesced around a few, large platform-based companies, such as Amazon, Microsoft and Alphabet, by the turn of the twenty-first century (Kenney & Zysman, 2016; Rahman & Thelen, 2019). Starting in the 2010s, those companies

pivoted away from traditional platform services to cloud computing and the supply of task-specific AI models that "could often be trained on a single server or a small cluster of servers" (Tan & Thelen, 2025, p. 14). Critically, however, those models have been replaced, over the last few years, by the use of "foundation models," of which large language models (LLMs) are the most well-known cases.

The development and use of foundation models entail a set of technological requirements that affect the market structure, capital concentration, and type of capital in the AI industry.

The AI model of production (through foundation models) consists of two main processes: model pretraining and the direct provision of (mostly inference) services. Pretraining, necessary to generate foundation models, requires massive and growing computational resources. Although unit inference costs are low (and quickly falling), aggregate inference costs generally exceed training costs over a model's operational life.

Since 2020, the computational power used to train frontier AI models has multiplied by a factor of five every year. Power usage to train those models has grown by a factor of 2.4 per year (Epoch AI, 2025). The fact that the performance of foundation models increases with the amount of compute used in training the model (Kaplan et al., 2020) implies that the demand for compute power is set to accelerate in the near future. The cost of training frontier AI models has grown by a factor of two to three times per year for the past eight years, suggesting that the largest models may cost over a billion dollars by 2027 (Epoch AI, 2025). The significant costs imposed by the computational needs of foundation models are reinforced by a limited supply of both high-quality training data and human talent.

7.4.2 Capital Concentration in the AI Industry

The combination of relatively low variable costs and high fixed costs generates large economies of scale in the AI industry. In addition, the latter enjoys substantial economies of scope because its current technologies can be deployed across many different economic sectors. Together, these economies of scale and scope foster the formation of a market with a shrinking number of firms and a tendency toward natural monopoly.

Microsoft, AWS and Google hold about three-fourths of the combined generative AI software, model management platforms, and services market, which skyrocketed from a mere $191 million market in 2022 to $25.6 billion in 2024 (IoT-Analytics, 2025).[11] In the AI hardware market and, specifically, data center Graphic Processing Units, which grew sharply from $17 billion in 2022 to $125 billion in 2024, NVIDIA accounts for over 90 percent of sales (Fernandez, 2025). As of October 2025, for example, private LLM developers with a valuation of $1 billion or more summed up to around $1 trillion.

[11] This market does not include the market for individual generative AI applications (such as ChatGPT).

PROOFS

The top four private LLM developers represented 90 percent of that valuation: OpenAI, xAI, Anthropic, and Databricks had valuations of $500, $200, $183, and over $100 billion, respectively.[12]

This trend toward market concentration has been reinforced by vertical integration (Korinek & Vipra, 2025). Upstream integration has consisted of large technology companies combining the design of chips and foundation models (e.g., Google or Amazon designing their own AI accelerator chips, or NVIDIA offering foundation models) and establishing exclusive contracts between cloud computing services, data-rich companies and LLM developers. Downstream integration has proceeded through the integrated use of foundation models among cognitive workers and in commercial applications.

7.4.3 Capital Concentration across the Economy

Technological change, particularly through AI and automation, seems to be contributing also to an increasing concentration of capital and capabilities among larger firms across the overall economy. Firms with greater cash reserves and R&D capacity are more likely to demand AI-related labor, apparently experiencing higher productivity gains and innovation outcomes as a result. Alekseeva et al. (2021) document a growing demand for AI skills between 2010 and 2019, with associated wage premiums and a concentration of AI adoption in larger, more innovative firms. Larger firms are much more likely to adopt automation technologies, contributing to labor substitution and accounting for 20–30 percent of the labor productivity gap between large companies and the median firm (Acemoglu et al., 2022). Artificial intelligence investment leads to higher sales, employment, and market valuation, particularly in concentrated industries. Artificial intelligence–adopting firms shift toward more educated STEM and IT workforces and flatten organizational hierarchies by reducing middle management roles (Babina et al., 2024). These changes amplify the advantages of large firms. Automation has stronger adverse effects on employment and wages in small firms and among older and middle-educated workers (Bessen et al., 2020).

7.4.4 Shifting Type of Capital

The shift of large IT companies toward the development of AI models has arguably cemented their current market dominance. At the same time, it appears to be transforming the nature of their assets. Platform companies were "asset light" (Tan & Thelen, 2025). Their product model and market position rested on the exploitation of networks of contracts, data-intensive operations, and highly educated personnel. Today, the demand for large agglomerations of computational power to pretrain and run LLMs requires the construction of large data centers and the consumption of vast amounts of energy. As a result,

[12] Data collected by the authors.

PROOFS

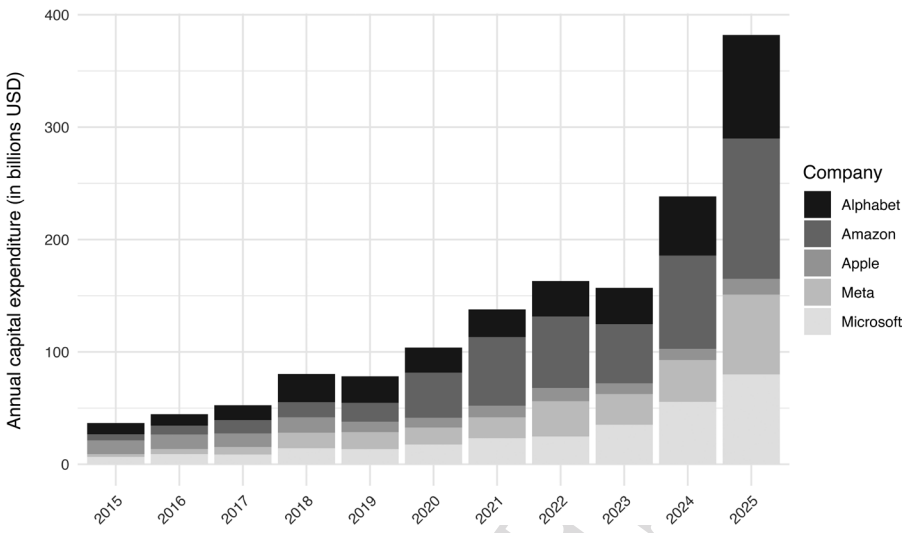


FIGURE 7.2 Capital expenditures of top five technological companies, 2015–25.
Note: Data collected by authors. Main source: FinanceCharts.

technology companies are becoming massive owners of (immobile) physical infrastructure.

As shown in Figure 7.2, capital expenditures by the five top companies grew from $36 billion in 2015 to around $150 billion in 2023. In the last two years, it has more than doubled to $400 billion. This represents over a third of all private fixed investment in information processing equipment and software.[13] Morgan Stanley projects $2.9 trillion in spending from 2025 to 2028 on chips, servers and data-center infrastructure.[14]

7.5 AI AND DEMOCRACY

7.5.1 The Electoral Arena

7.5.1.1 The Electoral Consequences of the IT and Globalization Shocks

Automation and digital technologies, compounded by the lack of response from most mainstream parties, reshaped the electoral arena of postindustrial democracies in the last few decades in the following way (rendered here in a simplified manner for space reasons).

[13] U.S. Bureau of Economic Analysis, Private fixed investment in information processing equipment and software [A679RC1Q027SBEA], retrieved from FRED, Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/A679RC1Q027SBEA.

[14] *Wall Street Journal*, July 31, 2025, https://tinyurl.com/wsj-aispend

PROOF

Rising wage inequality and rapid skill-biased employment change polarized voters' preferences. While highly educated voters continued to support the existing policy status, both white-collar voters and industrial workers with routine jobs and/or working in industries directly exposed to automation and foreign competition grew increasingly alienated from existing parties and their policy platforms.

That rising political alienation, combined with the limited response of mainstream parties – particularly left-wing parties that had traditionally represented low-income voters – to these economic transformations, triggered a broader process of electoral dealignment. In Western Europe, abstention grew from one out of five people in 1980 to one in three thirty years later, becoming highly and negatively correlated with income and age. Trust toward institutions and politicians plummeted.

Exploiting voters' dissatisfaction, a number of "populist" candidates jumped into the electoral arena, successfully inducing a partial process of electoral realignment (Boix, 2019). Higher exposure to robotization raised support for populist candidates in Europe (Anelli et al., 2019, 2021, Caselli et al., 2021, Milner, 2021) and the United States (Frey et al., 2018, Gonzalez-Rostani, 2025). Likewise, exposure to technological change increased support for illiberal preferences – such as trade protectionism and anti-immigration positions – commonly espoused by populists (Gonzalez-Rostani, 2024, Chaudoin & Mangini, 2025). In response to this populist rise, some left-wing parties have recently pivoted toward more expansive tax-and-spend platforms, in some cases partially embracing elements of the populist agenda, alongside more direct redistributive policies.

7.5.1.2 The Electoral Consequences of the AI Shock

The AI shock could eventually transform the current electoral market too. As with the political effects of globalization and the digital shock, the behavior and potential electoral realignment of voters will be a function of, first, changing economic conditions and, second, the response of political incumbents (and political entrepreneurs) to those changes.

Looking at the AI vulnerability index(es) we presented in Section 7.2 provides us with an entry point to measure the economic (and therefore potential electoral) impact of AI.

Figure 7.3 plots the median AI exposure level according to three popular measures by income decile in the United States. To that effect, we use the 2023 American Community Survey five-year file (a 5-in-100 national random sample of the population), match Census occupations to z-standardized AI exposure, and divide the sample into (weighted) deciles of respondents' pretax wage and salary income.[15] The plot shows box plots by decile for each

[15] The reference period for wage income is the previous twelve months. Income values are expressed in 2023 US dollars. The analysis sample consists of 3,850,246 matched cases.

PROOFS

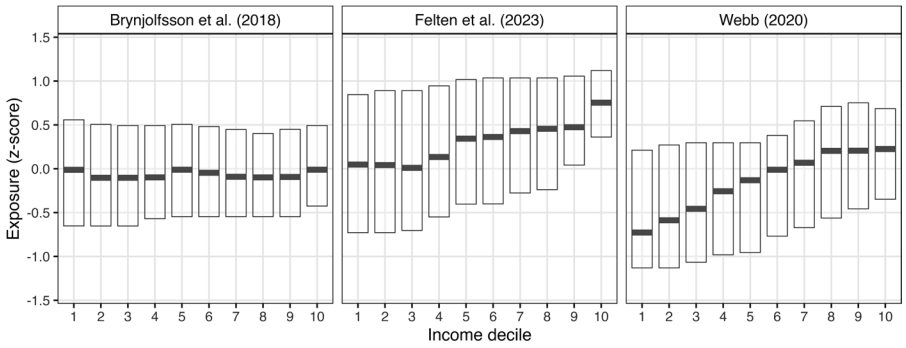


FIGURE 7.3 AI exposure by income decile in the US.

measure.[16] The Brynjolfsson exposure index hardly varies across the income distribution (we estimate its correlation with income as –0.01). The Felten index implies that high-income deciles are more vulnerable than the average (correlation with income: 0.23). The Webb index also indicates a correlation of income with AI vulnerability (estimated as 0.2) but shows that low-income deciles are less vulnerable than the mean.

As discussed earlier, digitalization and computerization eroded the position of occupations characterized by routine tasks and hollowed out middle social strata, raising wage inequality and allegedly polarizing politics. The Brynjolfsson et al. vulnerability index does not predict any relative change across the voters with respect to where they were after the digital and globalization shocks.

By contrast, the Felten and Webb indexes, which seem to forecast a wave of nonroutine-jobs substitution, entail a reversal of the recent trend toward more inequality. Mechanically, less wage inequality could have two effects: first, some depolarization in elections; second, a political realignment of capital owners and low-skill, manual occupations against high-skilled, AI-vulnerable employees.

Even if that economically equalizing and electorally depolarizing scenario takes place, a fall in the demand for high-skilled jobs could still exacerbate political tensions. Lower wages among highly educated individuals could trigger more redistributive conflict, ignited by a drop in relative status and frustrated social expectations among educated voters. In addition, if AI turns out to affect new entrants in the labor market more than older cohorts, intergenerational conflict may become central in elections.

So far we have pointed to AI's short- and medium-term electoral effects. In the long run, full automation (replacing non-manual jobs with embodied AI) and its correlated steep returns to capital could overrun those initial depolarization effects. It would also foster the emergence of a strong electoral cleavage between capital and labor (Boix, 2024).

[16] To make the plot visually cleaner, it shows only the medians and inter-quartile range box.

PROOFS

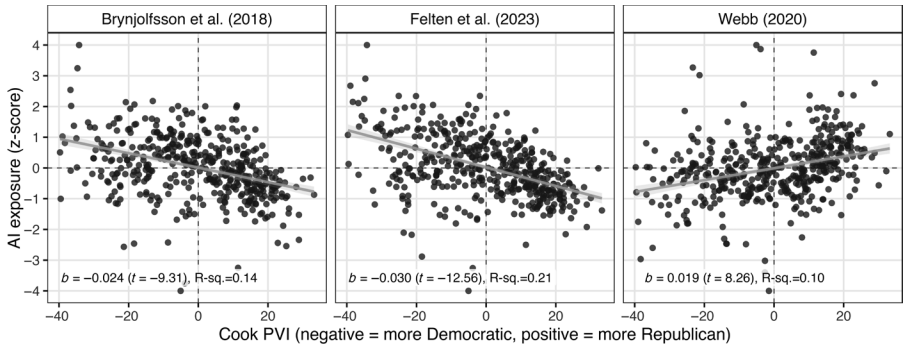


FIGURE 7.4 District-level AI exposure and partisan vote in the US.

At the end of the day, the electoral consequences of AI-driven economic change will be mediated by the strategic decisions of parties. Figure 7.4 plots the correlation between the three AI exposure indexes and district partisanship in the US. The vertical axis uses district-level exposure calculated by matching employment-weighted industry-level exposure to industry-level employment shares in each district of the 118th Congress, while the horizontal axis employs Cook's Partisan Voting Index, which captures the two-party vote margin compared to the national average. The beta coefficients noted in each panel show the linear relationship between the two.[17]

The level of AI vulnerability by partisan support varies considerably across indexes. According to the Brynjolfsson and Felten measures, Democratic-leaning districts are more vulnerable to AI than Republican-leaning ones. In the Webb index, Republicans are (moderately) more exposed to AI. In light of the first two indexes, Republicans may have a lower electoral incentive to oppose and regulate AI. This could explain the relative preponderance of pro-AGI effective accelerationism in the current federal government – something reinforced by its current strong connections with technology companies.[18]

7.5.2 Impact on Democratic Institutions

To understand the potential effects of AI on democratic institutions, consider a simple set-up in which voters vote for the candidate that offers a policy platform

[17] The coefficients are calculated from a robust (median) regression. Note that the vertical axes are clipped at four standard deviations for presentational purposes. The beta coefficients are based on all districts.

[18] Besides concerns about the employment and income effects of AI, public opinion is increasingly worried about the location of data centers and their impact on energy prices. As of November 2025, 41 percent of Americans favored a ban on the construction of data centers near where they live. About 36 percent opposed it. Support for a ban was marginally higher among Democrats. See Yokley (2025).

PROOFS

that maximizes their welfare. Each candidate's policy platform specifies a level of redistribution across voters through a combination of taxes, transfers and regulatory measures – where the latter refer to the legal rules governing capital, labor and goods markets (such as migration policies or tariffs that determine the supply of labor or goods).

Voters differ in their income, which is a function of their respective occupations (in turn, defined as a bundle of tasks). Voters' preferred level of redistribution is negatively related to their income. At the same time, because the marginal utility of income declines as individual income increases, the intensity with which voters may oppose redistributive policies falls as they become richer.

Because politicians are policy-oriented (and there is some uncertainty about the position of the median voter), they offer distinctive platforms that do not converge fully around the preferences of the median voter. In fact, their policy promises vary with the distribution of voters' preferences. As the latter become more polarized, redistribution platforms tend to diverge from each other (Beramendi et al., 2026).

Crucially, as political preferences become more (less) intense and/or extreme, voters are more (less) likely to tolerate undemocratic behavior that reduces the probability that the parties they oppose win elections (Graham & Svolik, 2020). Hence, as politics becomes more (less) polarized, incumbent politicians may have more (fewer) incentives to act anti-democratically. For a high level of polarization, democratic backsliding could even result in a breakdown of democracy (Beramendi et al., 2026).

In this set-up, the political impact of AI will depend to a large extent on whether it will intensify or attenuate economic inequality among voters.

7.5.2.1 Pro-democratic Effects of AI

Artificial intelligence's economic effects may consolidate democratic institutions through, at least, two mechanisms. In the first place, if AI equalizes the distribution of wages, particularly by displacing nonroutine jobs, partisan competition may become more centripetal and therefore less susceptible to anti-democratic candidates. In the second place, even if wage dispersion increases or does not change, higher income growth (due to the productivity gains of AI) may attenuate the polarizing effects of inequality. With marginal utility of income converging to zero for very high income levels, those individuals benefiting directly from AI (and future embodied AI) may be willing to tolerate higher levels of redistribution (for example, through robot taxes).

7.5.2.2 AI and Democratic Backsliding

Conversely, AI may trigger a process of de-democratization for two reasons. It could intensify political polarization. And it could foster the rise of powerful technological companies with strong incentives to capture regulatory agencies and bypass the democratic process. We discuss the latter mechanism in Section 7.5.3.

PROOFS

Political polarization may be the result of, at least, two main transformations. If embodied AI replaces manual occupations, employment and wages among low-educated workers will fall and income inequality will rise. Moreover, by accelerating capital-labor substitution, AI could compound the inequality effect generated by information and digital technologies in the last decades. At the end of 2025, the stock wealth of US households equaled close to three times their annual disposable income, or two and six times as much as in the early 2000s and 1980s, respectively. The top income decile holds around 70 percent of all financial assets (Federal Reserve Bank of St. Louis). As pointed out before, to the extent that those effects override any of AI's potential equalizing effects on the wage distribution, there should be higher polarization, now between capital-holders and the rest of the population, and a leftward shift in the candidates' position.[19]

At this point, it is unclear which forces will prevail for two reasons. First, AI's effects are still uncertain. Second, inequality and growth balance against each other in the model we lay out before. Beramendi et al. (2026) find, after examining the universe of all democratic spells from 1900 to 2017, that inequality has not trumped the pro-democracy effects of income in highly developed countries. Still, one could think of a scenario in which rising inequality could cancel out the "democratizing" effects of income growth. As we discuss in the conclusion, many of these outcomes may be, to a great extent, endogenous to politics. Policymakers could establish mechanisms to address the sources of inequality, such as investing in the generation of new skills among voters. They could also employ the wealth of today's advanced countries to compensate the losers of economic change.

7.5.3 Regulatory Capture

Artificial intelligence companies' market dominance and heavy investment in fixed assets, which make them more vulnerable to taxation, are likely to increase their opposition to regulatory measures to sustain market competition and break their dominant position.

Growing market dominance, which is a function of barriers to entry in the industry, makes businesses more concerned about policy. Any change in their input and output prices (potentially induced by government decisions) affects them more than if markets were highly competitive and new entrants could dissipate the benefits of preexisting firms (Frieden, 1991). Market dominance also has a direct effect on the collective action capacity and political clout of any industry. As the number of companies declines, the cost of lobbying governments falls. All else equal, market dominance reinforces the so-called structural power of capital over governments – the power that comes from the

[19] Since the holders of assets are older than average, political conflict would also turn around intergenerational distributive issues.

PROOFS

capacity to punish the latter with less investment if policy conditions are not favorable to potential investors.

The specificity of capital also affects the dependence of businesses with respect to governments. Owners of highly specific assets, that is, assets with value that drops substantially when put to an alternative use, have a strong incentive to shape policy because they can escape from governmental regulations only at a high cost (Boix, 2003, 2024). (This policy capture may include relying on public institutions to bail them out in the case of a systemic crash following current levels of investment.) In other words, firms substitute lobbying and capturing policymakers for the loss in the structural power that characterizes mobile capital.[20]

The market dominance of platform companies likely accounts for their strong current lobbying activity before American and European regulators.[21] The growing specificity of their current fixed capital investments likely encourages them to develop even stronger and more direct ties with policymakers (Tan & Thelen, 2025). In fact, it may explain why several large technology companies, shelving the liberal-leaning agenda they used to promote in the 2010s, are cultivating direct links with the current US administration.

7.6 ECONOMIC/POLITICAL BACKSLIDING ACROSS THE WORLD?

So far we have examined the potential impact of AI at the individual and sectoral level, mainly in the context of developed economies, which are capital-abundant and have a labor force of skilled and semi-skilled workers. Consider now the case of the Global South, fundamentally abundant in low-skilled labor.

Globalization benefited developing economies, making them deeply embedded in global value chains, and fostering a process of economic convergence with the North, particularly for East Asia and Eastern European and Latin American regions close to postindustrial countries. Artificial intelligence may reverse those trends. Still, because generative-AI is very recent – and cross-country evidence on adoption and realized effects is thin – claims about international spillovers are necessarily more speculative than what we can say about earlier automation and about domestic effects where the technology is directly deployed.

On the one hand, AI-driven changes in production, sourcing, and coordination can have positive or neutral effects in the Global South. East and South Asian economies are developing technological hubs or technology-intensive service platforms. Artificial intelligence may foster new economic opportunities,

[20] On current debates and strategies to regulate data centers, see Brennen and Mohammed (2025).
[21] See Halpin and Nownes (2021) for the United States and Gorwa et al. (2024) for the European Union.

PROOFS

including expanded remote and freelance services (Baldwin & Forslid, 2023). On the other hand, AI may have a negative impact on developing economies. Access to AI tools that could augment work is likely to be uneven due to important supply constraints: reliable connectivity and electricity, data infrastructure, and AI-relevant skills vary widely, shaping whether AI complements workers or substitutes for them (Cazzaniga et al., 2024; Gmyrek et al., 2023). Perhaps more importantly, AI may reduce the costs of capital relative to labor, accelerating production reshoring to the North.

For now, the clearest evidence on cross-border transmission comes from a pre-generative-AI and easier-to-measure technology: industrial robots. Automation costs have fallen sharply in the last decades. Robot prices declined by roughly 80 percent between 1995 and 2017 (Potter, 2024), making foreign capital far more likely to replace labor.

The offshoring boom of the early 2000s slowed down as automation made low wages less decisive for many tasks (Antràs, 2020; Faber et al., 2023). Growing research associates automation with less offshoring and more reshoring across firms and countries (De Backer et al., 2018; Krenz et al., 2021; Pinheiro et al., 2023; Bonfiglioli et al., 2024; Hidalgo & Micco, 2024). Those developments may affect economies specialized in labor-intensive manufacturing in particular (Rodrik, 2018; Stemmler, 2023). Studies linking US robotization to outcomes in partner countries find losses centered in export-oriented, routine manufacturing sectors, with stronger impacts in emerging economies than in richer ones (Carbonero et al., 2020; Faber, 2020; Kugler et al., 2020; Díaz Pavez & Martínez-Zarzoso, 2024). Distributional costs often fall on workers with weaker outside options, including many women, older workers, and low-educated production workers (Faber, 2020; Kugler et al., 2020).

Nonetheless, trade effects are not one-directional. Productivity gains in advanced economies can also increase demand for some imported inputs, including commodities, such as mining inputs, contributing to greater reliance on resource-based imports from the Global South (Stemmler, 2023).

The World Bank estimates that roughly 1.8 billion jobs in developing countries – about two-thirds of their workforce – are potentially automatable (Frey & Rahbari, 2016). If anything, AI could extend this logic beyond manufacturing. Firms could increasingly use AI to deliver services that were previously offshored – from customer support and back-office processing to parts of software development, accounting, translation, and other higher skill professional services. If that substitution scales, developing economies could face external AI shocks even in sectors where they once held a clear advantage in tradable services, and even where local AI adoption has started to take place.

While direct evidence on AI is still scarce, existing work on pre-generative-AI automation suggests that technological change in the Global North can have substantial political spillovers in the Global South. Boix et al. (2025), who study exposure to foreign robot adoption across Mexican commuting zones, find that areas more exposed to foreign robots shift leftward

PROOFS

politically – contrasting with evidence from advanced economies, where similar shocks often fuel right-wing populism. The authors attribute this difference to the structure of political coalitions in labor-abundant economies. Low-skill workers historically benefited from globalization, but foreign automation and reshoring threaten these gains, weakening pro-market coalitions and increasing support for redistribution and protection. The same exposure is also associated with higher levels of organized crime, consistent with broader evidence that adverse labor-market shocks can increase violence when formal employment opportunities contract (Dube & Vargas, 2013; Dix-Carneiro et al., 2018; Dell et al., 2019).

Artificial intelligence's long-run impact on political institutions may be even more dramatic. If it results in high levels of automation and reshoring, industrializing countries may be unable to escape from a middle-income trap. In the limit, they may even experience negative growth rates and a shrinking middle class and, as result, democratic backsliding or, in those places without free elections, the strengthening of authoritarian institutions (Boix, 2024).

Low fertility rates and aging in the advanced world account for part of recent South-North migratory flows. If at some point AI and robotics substitute for some migrant labor in care, agriculture, and other services, they will dampen Global South out-migration, increasing labor-absorption pressures at home.

7.7 CONCLUSIONS

In this chapter, we have examined the political economy of AI-driven transformation in labor markets, capital, and the production process in general.

In line with reports such as Narayanan & Kapoor (2025), we tend to view AI as a "normal" (general-purpose) technology, at least at this point in time and in regard to labor markets. First, the introduction of AI appears to have been gradual and circumscribed to a fraction of the multiple tasks that generally compose any occupation. By replacing specific tasks, AI is arguably allowing its direct users to expand their attention to the other components of their specific occupation. Second, the effects of AI appear to be conditional on its complementary with the skills of its users and mediated by the institutional environment in which it is diffused. Perhaps for these reasons, the public holds ambiguous views about the social and economic consequences of AI, making elite discourse central to its reception and regulation.

Artificial intelligence may be having a more substantial (but, so far, less researched) impact on capital. The economics of foundation models have reinforced market concentration in the technological sector and triggered massive investments in fixed assets. This, in turn, has boosted the lobbying capacity of technological firms and intensified their appetite for regulatory capture.

The disruptive effects of AI in the labor market and the growing role played by tech capitalists could exacerbate the political tensions that have recently characterized advanced democracies. At the same time, however,

PROOFS

an AI-generated positive productivity shock may attenuate electoral conflict. Overall, these effects will be conditional on political responses. For example, stepping up AI-related training of the labor force could smooth the transition to a new production paradigm.

As stressed throughout this chapter, we still know little about the economic and political consequences of AI. This is undoubtedly due to the short period elapsed since the practical application of AI models to economic tasks. But part of it seems due to important disagreements on how to measure its effects and to the difficulty of translating what we know in terms of AI methods to their actual application in specific tasks and occupations. Hence, a critical line of research should consist of understanding the ways in which jobs are exposed to AI – and of reconciling the different vulnerability or exposure to AI measures that have been developed so far.

Future research on the political and social effects of AI should probably avoid any deterministic understanding of its impact. First, the extent to which AI will reshape society will depend on the speed of adoption (gradual, fast and massive, or even in the form of a technological singularity) as well as on our capacity for adjustment (Milner, 2025). Second, its regulatory environment, and therefore its impact, will be a function of the discursive framing of political entrepreneurs and legal decisions made by policymakers. Finally, its political effects will vary with the policy mechanisms developed to compensate the losers of change and to share the benefits of these new technologies.

When designing a policy tool kit in response to AI's disruptive effects, we need to consider two diverging future scenarios. In the first one, AI intensifies (or even reverses) the skilled-biased nature of current technological changes. In the second one, AI (mainly embodied AI) substitutes for most or all labor. Particularly in the second scenario, capital would represent a disproportionate share of national income.

The urgency and reach of the three main policy interventions we outline further next will depend on which of the two scenarios we end up facing and on the extent of the disruption generated by AI. First, investing in education (including vocational training) should make the workforce easier to match to AI-complementary jobs. If the process of capital-labor substitution generates some persistent unemployment or permanently falling wages, countries could set a "universal basic capital" (UBC) system, giving everyone some fixed capital at birth. To incentivize prudential behavior, free disposition over its returns would happen at adult age. Universal basic capitals resemble the idea of a universal basic income (UBI) but give economic agency to their holders. Universal basic capitals could be funded through a tax on robots and structured as pension funds operate today. One may think of this policy as a system to maintain markets while constructing a society of capital holders. An alternative (or complementary) intervention could consist of creating a sovereign fund, funded through taxes or through the returns to shares in capital-intensive companies, that would sustain social spending.

PROOFS

Second, democratic institutions should invest in creating a legal framework that blocks the formation of a closed elite. That would require the combination of an active antitrust policy and of measures to reinforce political participation, reduce unequal electoral funding, publicize (ownership and marketing) relations between media and large firms, strengthen local level governance schemes, and so on.

Finally, although the economic and political repercussions of AI-driven automation for the Global South remain uncertain, they warrant closer political scrutiny. Policymakers may want to give some consideration to interventions that mitigate AI's potential negative effects on growth (through reshoring) there. In a world of stagnating or even declining population, advanced economies (or at least those clusters experiencing AI-productivity shocks) may wish to be open to immigration – both as a mechanism to equalize life chances across the world and to sustain welfare states in the North.

APPENDIX

TABLE 7.A *Overview of task and automation exposure objective measures*

Reference	Measure	What it captures
(Autor et al., 2003)	Task groups	Canonical split of jobs into routine versus nonroutine analytical, interactive, or manual. Relied on O*NET task data to classify occupational activities into three categories: routine tasks – rule-based activities that can be easily automated (e.g., record-keeping, assembly). Nonroutine cognitive tasks – analytical or problem-solving work requiring adaptability and judgment. Nonroutine manual tasks – physical tasks demanding dexterity or situational flexibility that are difficult to automate. This classification forms the foundation of the task-based framework for understanding how technology reshapes labor demand.
(Goos et al., 2014)	Routine Task Intensity (RML)	An occupation-level index built from five DOT task variables (following Autor–Levy–Murnane). Tasks are grouped into Routine, Abstract, and Manual, and RTI is computed as $RTI = ln(Routine) - ln(Abstract) - ln(Manua)$. Industry- and country-level RTI for Europe by mapping occupational task content (O*NET/DOT to ISCO) and aggregating to industries; used to explain polarization and the role of offshoring.

(*continued*)

PROOFS

TABLE 7.A (*continued*)

Reference	Measure	What it captures
(Frey & Osborne, 2017)	Computerization probability	Occupation-level automatability from expert-labeled training set and a classifier using O*NET features; yields the probability a job is susceptible to computerization.
(Arntz et al., 2017)	Task-based automation risk	Worker-level automation probabilities using Program for the International Assessment of Adult Competencies task surveys; accounts for within-occupation task heterogeneity and task bundling when estimating risk.
(Graetz & Michaels, 2018)	Robot intensity	Industry-level stock/flow of industrial robots with data from International Federation of Robotics (IFR) in particular "multipurpose manipulating industrial robots." The stock was normalized by hours worked across countries; proxy for exposure to robot adoption in production.
(Brynjolfsson et al., 2018)	Suitability for Machine Learning (SML)	Suitability for machine learning scores: task-level ratings on 23 criteria (e.g., data availability, definable objectives), aggregated to occupations to flag ML-amenable tasks. These ratings were applied to the textual descriptions of all O*NET occupations via CrowdFlower, a crowdsourcing platform. The resulting index provides a forward-looking estimate of which tasks are most likely to be automated through machine learning or AI in the near future.
(Felten et al., 2019)	AI Occupational Exposure (AIOE)	Uses data from the Electronic Frontier Foundation's AI Progress Measurement project. It links nine AI application areas (e.g., image recognition, language modeling) to fifty-two O*NET ability scales through crowd-sourced relevance scores from Amazon MTurk. Each occupation's AI exposure is computed as a weighted sum of these application-ability scores, with weights reflecting the importance and prevalence of each ability in that occupation.
(Acemoglu & Restrepo, 2020)	Robot exposure	Commuting-zone exposure to robots: IFR industry robot penetration weighted by local 1990 industry employment shares; links robot diffusion to local labor outcomes.
(Webb, 2020)	AI patent exposure	Textual overlap between job task descriptions (O*NET) and AI patent text to measure direct exposure of occupations to AI technologies (also constructed for software/robots).

(*continued*)

PROOFS

TABLE 7.A (*continued*)

Reference	Measure	What it captures
(Anelli et al., 2021)	Vulnerability to automation	Individual-level exposure combining (1) predicted occupation probabilities based on pretreatment characteristics (age, gender, education); (2) occupation automatability/robot relevance; and (3) country–year robot adoption to gauge personal vulnerability.
(Manning et al., 2023)	LLM exposure (GPT-4)	Task- and occupation-level exposure based on alignment between O*NET tasks and LLM capabilities, using expert judgments and GPT-4 assessments; higher values indicate greater potential task impact.
(Pizzinelli et al., 2023)	AI exposure & complementarity	Extension of exposure metrics that separates exposure from complementarity by incorporating work context and skills; distinguishes high-exposure jobs where AI is more likely to complement versus substitute labor.
(Kogan et al., 2023)	Labor-saving versus labor-augmenting exposure (patent–task)	Text similarity between patents and occupation tasks to build time-varying exposure measures that distinguish substitution (labor-saving) from augmentation (labor-augmenting) at worker/occupation levels. Patents whose language closely matches routine tasks are treated as labor-saving (automation); matches to nonroutine, skill-intensive tasks are treated as labor-augmenting (complementary).
(Eisfeldt et al., 2023)	Generative AI workforce exposure	Firm-level exposure derived by mapping generative-AI-susceptible occupational tasks to each firm's workforce mix; used to study market valuation effects of GenAI.
(Agnolin & González-Rostani, 2025)	monitoring, augmentation	(1) Map ISCO occupations to six families of workplace technologies using the TechXposure patent-text embedding method (machine learning, embedded systems, remote monitoring, smart mobility, intelligent logistics, food-ordering), then scale each occupation's score by the respondent's country-year Enterprise Resource Planning systems (ERP) adoption rate from Eurostat to capture diffusion, producing individual-level exposure. (2) Build an LLM-based index at the industry-occupation level that rates augmentation, monitoring, and replacement on 1–10 scales with paired certainty scores and binary exposure flags.

(*continued*)

PROOFS

TABLE 7.A (*continued*)

Reference	Measure	What it captures
(Prytkova et al., 2025)	Emerging technologies and patents	Measured by matching patent text to standardized industry (NACE Rev.2, three-digit) and occupation (ISCO-08, four-digit) descriptions using sentence-transformer embeddings (all-mpnet-base-v2) and cosine similarity. These similarities are then aggregated within forty technology clusters obtained by k-means on patent embeddings, weighted by patent citations, and transformed with an inverse-hyperbolic-sine to reduce skew, yielding technology-specific exposure scores for each industry and occupation.
(Tomlinson et al., 2025)	AI applicability	Score built from real-world usage of a generative-AI assistant, by classifying Bing Copilot conversations into work activities (e.g., gathering information, writing, advising) and recording how often and how successfully AI performs each. They then map those activities to the activity mix of occupations, weighted by measured task success and breadth of impact, to produce a per-occupation score capturing how much current generative-AI can be productively applied to that job.

PROOFS

REFERENCES

Acemoglu, D., Aghion, P., & Violante, G. L. (2001). Deunionization, technical change and inequality. *Carnegie-Rochester Conference Series on Public Policy*, *55*(1), 229–264.

Acemoglu, D., & Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. In Orley Ashenfelter, & David Card (Eds.), *Handbook of labor economics* (Vol. 4, pp. 1043–1171). Elsevier.

Acemoglu, D., Autor, D., Hazell, J., & Restrepo, P. (2022). Artificial intelligence and jobs: Evidence from online vacancies. *Journal of Labor Economics*, *40*(S1), S293–S340.

Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, *33*(2), 3–30.

Acemoglu, D., & Restrepo, P. (2020). Robots and jobs: Evidence from US labor markets. *Journal of Political Economy*, *128*, 2188–2244.

Acemoglu, D., & Restrepo, P. (2022). Tasks, automation, and the rise in U.S. wage inequality. *Econometrica*, *90*(5), 1973–2016.

Acemoglu, D., & Restrepo, P. (2024). *Automation and rent dissipation: Implications for wages, inequality, and productivity*. No. 32536. National Bureau of Economic Research, Inc.

Agarwal, N., Moehring, A., Rajpurkar, P., & Salz, T. (2023). *Combining human expertise with artificial intelligence: Experimental evidence from radiology* (No. w31422). National Bureau of Economic Research.

Agnolin, P., Anelli, M., Colantone, I., & Stanig, P. (2025). *Robots replacing trade unions: Novel data and evidence from western Europe*. Institute of Labor Economics (IZA) Discussion Papers, No. 17864, Bonn.

Agnolin, P., & González-Rostani, V. (2025). *When technology manages: Workers' demands and union responses to AI and emerging digital tools*. Working Paper. University of Southern California.

Ahlquist, J. (2017). Labor unions, political representation, and economic inequality. *Annual Review of Political Science*, *17*, 409–432.

Alekseeva, L., Azar, J., Giné, M., Samila, S., & Taska, B. (2021). The demand for AI skills in the labor market. *Labour Economics*, *71*, 102002.

Anelli, M., Colantone, I., Gallego, A., & Stanig, P. (2025). What if you see it? Workers' perceptions of and reactions to LLMs. Working Paper. Bocconi University.

Anelli, M., Colantone, I., & Stanig, P. (2019). *We were the robots: Automation and voting behavior in Western Europe* (No. 17/19). Centre for Research & Analysis of Migration (CReAM) Discussion Paper Series, Department of Economics, University College London, London.

Anelli, M., Colantone, I., & Stanig, P. (2021). Individual vulnerability to industrial robot adoption increases support for the radical right. *Proceedings of the National Academy of Sciences*, *118*(47), e2111611118.

Antràs, P. (2020). *De-globalisation? Global value chains in the post-COVID-19 age* (No. w28115). National Bureau of Economic Research.

Arntz, M., Gregory, T., & Zierahn, U. (2017). Revisiting the risk of automation. *Economics Letters*, *159*, 157–160.

Autor, D. (2013). The "task approach" to labor markets: An overview. *Journal for Labour Market Research*, *46*(3), 185–199.

Autor, D. (2024). *Applying AI to rebuild middle class jobs* (No. w32140). National Bureau of Economic Research.

Autor, D. H., Katz, L. F., & Krueger, A. B. (1998). Computing inequality: Have computers changed the labor market? *The Quarterly Journal of Economics*, *113*(4), 1169–1213.

Autor, D. H., Levy, F., & Murnane, R. J. (2002). Upstairs, downstairs: Computers and skills on two floors of a large bank. *ILR Review*, *55*(3), 432–447.

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics*, *118*(4), 1279–1333.

Autor, D., & Salomons, A. (2018). *Is automation labor-displacing? Productivity growth, employment, and the labor share* (No. w24871). National Bureau of Economic Research.

Babina, T., Fedyk, A., He, A. X., & Hodson, J. (2023). *Firm investments in artificial intelligence technologies and changes in workforce composition* (Vol. 31325). National Bureau of Economic Research.

PROOFS

Babina, T., Fedyk, A., He, A., & Hodson, J. (2024). Artificial intelligence, firm growth, and product innovation. *Journal of Financial Economics*, *151*, 103745.

Balcázar, C. F. (2023). *Globalization, unions and robots: The effects of automation on the power of labor and policymaking*. Available at SSRN 4574527: https://papers.ssrn.com/sol3/Delivery.cfm?abstractid=4574527.

Baldwin, R., & Forslid, R. (2023). Globotics and development: When manufacturing is jobless and services are tradeable. *World Trade Review*, *22*(3–4), 302–311.

Becher, M., & Stegmueller, D. (2021). Reducing unequal representation: The impact of labor unions on legislative responsiveness in the U.S. congress. *Perspectives on Politics*, *19*(1), 92–109.

Becher, M., & Stegmueller, D. (2025). *Machines against workers? The heterogeneous impact of robots on union strength*. Working paper. Instituto de Empresa and Duke University.

Becker, Joel, Rush, Nate, Barnes, Elizabeth, & Rein, David. (2025). Measuring the impact of early-2025 AI on experienced open-source developer productivity. *arXiv preprint arXiv:2507.09089*.

Beramendi, P., Boix, C., & Stegmueller, D. (2026). Resilient democracies. *Journal of Politics*. https://doi.org/10.1086/736562.

Bessen, James, Goos, Maarten, Salomons, Anna, & Van den Berge, Wiljan. (2025). Automatic reaction-what happens to workers at firms that automate? *The Review of Economics and Statistics*, 107(1), 125.

Bessen, J., Goos, M., Salomons, A., & Berge, W. van den. (2020). *Automation: A guide for policymakers*. Washington, DC: Economic Studies at Brookings Institution

Boix, C. (2003). *Democracy and redistribution*. Cambridge University Press.

Boix, C. (2019). *Democratic capitalism at the crossroads*. Princeton University Press.

Boix, C. (2024). AI and the economic and informational foundations of democracy. In Justin B. Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew M. Young, and Baobao Zhang (Eds.), *The oxford handbook of AI governance*. Oxford University Press.

Boix, C., Gonzalez-Rostani, V., & Owen, E. (2025). The political economy of automation and fragmented production: Evidence from Mexico. Available at SSRN: https://papers.ssrn.com/sol3/Delivery.cfm?abstractid=5553058.

Bonfiglioli, A., Crino, R., Fadinger, H., & Gancia, G. (2024). Robot imports and firm-level outcomes. *The Economic Journal*. *134*(664), 3428–3444.

Borwein, S., Bonikowski, B., Loewen, P., Magistro, B., & Lee-Whiting, B. (2024). Who can assert ownership over automation? Workplace technological change, populist and ethno-nationalist rhetoric, and candidate support. *Political Behavior*, *46*, 2191–2214.

Braxton, J. C., & Taska, B. (2023). Technological change and the consequences of job loss. *American Economic Review*, *113*(2), 279–316.

Brennen, Scott Babwah, & Mohammed, Lama. (2025). *The hidden regulators: Public utility commissions and AI governance*. NYU's Center on Technology Policy. November. https://tinyurl.com/5n6jcvas

Bresnahan, T. F., & Trajtenberg, M. (1995). General purpose technologies: "Engines of growth." *Journal of Econometrics*, *65*(1), 83–108.

Brynjolfsson, E., Chandar, B., & Chen, R. (2025). *Canaries in the coal mine? Six facts about the recent employment effects of artificial intelligence*. Stanford Digital Economy Lab. Published August https://tinyurl.com/rwncpf2s.

PROOFS

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics*, *140*, 889–942.
Brynjolfsson, E., Mitchell, T., & Rock, D. (2018). What can machines learn, and what does it mean for occupations and the economy? *AEA Papers and Proceedings*, *108*, 43–47.
Brynjolfsson, E., Rock, D., & Syverson, C. (2021). The productivity J-curve: How intangibles complement general purpose technologies. *American Economic Journal: Macroeconomics*, *13*(1), 333–372.
Burn-Murdoch, John, & O'Connor, Sarah. (2026a). The AI Shift: What millions of job postings are telling us. *Financial Times*. January 22. https://tinyurl.com/4yu5uk9t.
Burn-Murdoch, John, & O'Connor, Sarah. (2026b). The AI shift: Is this the "take off" moment for AI agents? *Financial Times*. February 5. https://tinyurl.com/5ahan7cu.
Busemeyer, M. R., Gandenberger, M., Knotz, C., & Tober, T. (2023). Preferred policy responses to technological change: Survey evidence from OECD countries. *Socio-Economic Review*, *21*(1), 593–615.
Carbonero, F., Ernst, E., & Weber, E. (2020). Robots worldwide: The impact of automation on employment and trade. Beiträge zur Jahrestagung des Vereins für Socialpolitik 2020: Gender Economics, ZBW – Leibniz Information Centre for Economics, Kiel, Hamburg.
Caselli, M., Fracasso, A., & Traverso, S. (2021). Globalization, robotization, and electoral outcomes: Evidence from spatial regressions for Italy. *Journal of Regional Science*, *61*(1), 86–111.
Cazzaniga, M., Jaumotte, M. F., Li, L., et al. (2024). *Gen-AI: Artificial Intelligence and the future of work*. International Monetary Fund.
Chaudoin, S., & Mangini, M.-D. (2025). Robots, foreigners, and foreign robots: Policy responses to automation and trade. *The Journal of Politics*. www.journals.uchicago.edu/doi/epdf/10.1086/737849.
Chiacchio, F., Petropoulos, G., & Pichler, D. (2018). *The impact of industrial robots on EU employment and wages: A local labour market approach* (No. 2018/02). Bruegel Working Paper.
Copestake, A., Marczinek, M., Pople, A., & Stapleton, K. (2023). *AI and services-led growth: Evidence from indian job adverts*. World Bank Policy Research Working Paper. March 6.
Dauth, W., Findeisen, S., Suedekum, J., & Woessner, N. (2021). The adjustment of labor markets to robots. *Journal of the European Economic Association*, *19*(6), 3104–3153.
De Backer, K., DeStefano, T., Menon, C., & Suh, J. R. (2018). *Industrial robotics and the global organisation of production*. OECD Science, Technology and Industry Working Papers 2018/03.
De Stefano, V., & Doellgast, V. (2023). Introduction to the transfer special issue. Regulating AI at work: Labour relations, automation, and algorithmic management. *Transfer: European Review of Labour and Research*, *29*(1), 9–20.
Dell, M., Feigenberg, B., & Teshima, K. (2019). The violent consequences of trade-induced worker displacement in Mexico. *American Economic Review: Insights*, *1*(1), 43–58.
Dix-Carneiro, R., Soares, R. R., & Ulyssea, G. (2018). Economic shocks and crime: Evidence from the brazilian trade liberalization. *American Economic Journal: Applied Economics*, *10*(4), 158–195.

PROOFS

Díaz Pavez, L. R., & Martínez-Zarzoso, I. (2024). The impact of automation on labour market outcomes in emerging countries. *The World Economy*, *47*(1), 298–331.
Doellgast, V., Appalla, S., Ginzburg, D., Kim, J., & Li Thian, W. (2025). *Global case studies of social dialogue on AI and algorithmic management* (Working Paper No. 144). International Labour Organization.
Dube, O., & Vargas, J. F. (2013). Commodity price shocks and civil conflict: Evidence from Colombia. *The Review of Economic Studies*, *80*(4), 1384–1421.
Eisfeldt, A. L., Schubert, G., Zhang, M. B., & Taska, B. (2023). *Generative AI and firm values* (No. w31222). National Bureau of Economic Research.
Ellingrud, K., Sanghvi, S., Madgavkar, A., et al. (2023). *Generative AI and the future of work in America*. McKinsey Global Institute. July.
Epoch AI. (2025). Key trends and figures in machine learning. https://epoch.ai/trends. Accessed on November 18, 2025.
European Commission. (2024). *Special Eurobarometer 554: Artificial intelligence and the future of work*. Publications Office of the European Union, Luxembourg.
Faber, M. (2020). Robots and reshoring: Evidence from Mexican labor markets. *Journal of International Economics*, *127*, 103384.
Faber, M., Kilic, K., Kozliakov, G., & Marin, D. (2023). Global value chains in a world of uncertainty and automation. Available at SSRN 4661011.
Farber, H. S., Herbst, D., Kuziemko, I., & Naidu, S. (2021). Unions and inequality over the twentieth century: New evidence from survey data. *Quarterly Journal of Economics*, *136*(3), 1325–1138.
Felten, E. W., Raj, M., & Seamans, R. (2019). The occupational impact of artificial intelligence: Labor, skills, and polarization. *NYU Stern School of Business*. SSRN. https://papers.ssrn.com/sol3/Delivery.cfm?abstractid=3368605.
Felten, E., Raj, M., & Seamans, R. (2021). Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. *Strategic Management Journal*, *42*(12), 2195–2217.
Felten, Edward W., Raj, Manav, & Seamans, Robert. (2023). How will language modelers like ChatGPT affect occupations and industries? *arXiv:2303.01157*.
Fernandez, J. (2025). *The leading generative companies*. IoT-Analytics. https://iot-analytics.com/leading-generative-ai-companies/.
Frey, C. B., Berger, T., & Chen, C. (2018). Political machinery: Did robots swing the 2016 US presidential election? *Oxford Review of Economic Policy*, *34*(3), 418–442.
Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, *114*, 254–280.
Frey, C. B., & Rahbari, E. (2016). *Do labor-saving technologies spell the death of jobs in the developing world?* Prepared for the 2016 Brookings Blum Roundtable. www.brookings.edu/wp-content/uploads/2016/07/Global_20160720_Blum_FreyRahbari.pdf.
Frieden, J. A. (1991). *Debt, development, and democracy: Modern political economy and latin america, 1965–1985*. Princeton University Press.
Friis, Simon, & Riley, James W. (2025). Performance or Principle: Resistance to Artificial Intelligence in the U.S. Labor Market. Harvard Business School Working Paper, No. 26–017, October 2025.

PROOFS

Gallego, A., & Kurer, T. (2022). Automation, digitalization, and artificial intelligence in the workplace: Implications for political behavior. *Annual Review of Political Science*, *25*(1), 463–484.
Gmyrek, P., Berg, J., & Bescond, D. (2023). *Generative AI and Jobs: A global analysis of potential effects on job quantity and quality*. ILO Working Paper 96 (Geneva, ILO). https://doi.org/10.54394/FHEM8239.
Golden, M. (1997). *Heroic defeats: The politics of job loss*. Cambridge University Press.
Gonzalez-Rostani, V. (2024). *The path from automation to populist political behavior*. Working Paper, University of Southern California.
Gonzalez-Rostani, V. (2025). Elections, right-wing populism, and political-economic polarization: The role of institutions and political outsiders. *The Journal of Politics*, 88(1).
Goos, M., & Manning, A. (2007). Lousy and lovely jobs: The rising polarization of work in Britain. *The Review of Economics and Statistics*, *89*(1), 118–133.
Goos, M., Manning, A., & Salomons, A. (2009). Job polarization in Europe. *American Economic Review*, *99*(2), 58–63.
Goos, M., Manning, A., & Salomons, A. (2014). Explaining job polarization: Routine-biased technological change and offshoring. *American Economic Review*, *104*(8), 2509–2526.
Gorwa, R., Lechowski, G., & Schneiß, D. (2024). Platform lobbying: Policy influence strategies and the EU's digital services act. *Internet Policy Review*, *13*(2), 1–26.
Graetz, G., & Michaels, G. (2018). Robots at work. *The Review of Economics and Statistics*, *100*(5), 753–768.
Graham, M. H., & Svolik, M. W. (2020). Democracy in America? Partisanship, polarization, and the robustness of support for democracy in the United States. *American Political Science Review*, *114*(2), 392–409.
Green, J., Grant, Z., Evans, G., & Inglese, G. (2025). Linking artificial intelligence job exposure to expectations: Understanding AI losers, winners, and their political preferences. *Research & Politics*, *12*(2), 20531680251337897.
Hall, P. A., & Soskice, D. (2001). *Varieties of capitalism: The institutional foundations of comparative advantage*. Oxford University Press.
Halpin, D., & Nownes, A. J. (2021). *The new entrepreneurial advocacy: Silicon valley elites in American politics*. Oxford University Press.
Haslberger, M., Gingrich, J., & Bhatia, J. (2025). Rage against the machine? Generative AI exposure, subjective risk, and policy preferences. *Journal of European Public Policy Preprint*.
Heinrich, T., & Witko, C. (2025). Self-interest and preferences for the regulation of artificial intelligence. *Journal of Information Technology & Politics*, *22*(3), 306–321.
Hertel-Fernandez, A. (2025). Civic organizations and the political participation of cross-pressured Americans: The case of the labor movement. *American Political Science Review*, *119*(3), 1173–1189.
Hidalgo, C., & Micco, A. (2024). Computerization, offshoring and trade: The effect on developing countries. *World Development*, *180*, 106617.
Hoffmann, F., Lee, D. S., & Lemieux, T. (2020). Growing income inequality in the United States and other advanced economies. *Journal of Economic Perspectives*, *34*(4), 52–78.

PROOFS

Huang, Y. (2025). *The labor market impact of artificial intelligence: Evidence from US regions*. IMF Working Paper, 24/199.
Humlum, A. (2019). *Robot adoption and labor market dynamics*. Rockwool Foundation Research Unit.
IOP. (2025). *Harvard youth poll* (51st ed.). Institute of Politics at Harvard Kennedy School.
IoT Analytics. (2025). Generative AI market report 2025–2030. https://iot-analytics.com/product/generative-ai-market-report-2025-2030/.
Iversen, T. (1999). *Contested economic institutions: The politics of macroeconomics and wage bargaining in advanced democracies*. Cambridge University Press.
Jäger, S., Naidu, S., & Schoefer, B. (2025). Chapter 4 – collective bargaining, unions, and the wage structure: An international perspective. In C. Dustmann & T. Lemieux (Eds.), *Handbook of labor economics* (Vol. 6, pp. 229–372). Elsevier.
Jaimovich, N., & Siu, H. E. (2019). How automation and other forms of IT affect the middle class: Assessing the estimates. *Brookings Economic Studies, Report*, November 2019, pp. 1–16.
Jeffrey, K. (2021). Automation and the future of work: How rhetoric shapes the response in policy preferences. *Journal of Economic Behavior & Organization*, *192*, 417–433.
Kaplan, E., & Naidu, S. (2025). Between government and market: The political economics of labor unions. *Annual Review of Economics*, *17*, 367–396.
Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). Scaling laws for neural language models. *arXiv:2001.08361*.
Karabarbounis, L. (2024). Perspectives on the labor share. *Journal of Economic Perspectives*, *38*(2), 107–136.
Kenney, M., & Zysman, J. (2016). The rise of the platform economy. *Issues in Science and Technology*, *32*(3), 61.
Kerrissey, J., & Schofer, E. (2013). Union membership and political participation in the united states. *Social Forces*, *91*(3), 895–928.
Kinder, M., Souza Briggs, X. de, Liu, S., & Muro, M. (2024). *Generative AI, the American worker, and the future of work*. Brookings. www.brookings.edu/articles/generative-ai-the-american-worker-and-the-future-of-work/, Zugriff, 24.
Koch, M., Manuylov, I., & Smolka, M. (2021). Robots and firms. *The Economic Journal*, *131*(638), 2553–2584.
Kogan, L., Papanikolaou, D., Schmidt, L. D., & Seegmiller, B. (2023). Technology and labor displacement: Evidence from linking patents with worker-level data (No. w31846). National Bureau of Economic Research.
Korinek, A., & Vipra, J. (2025). Concentrating intelligence: Scaling and market structure in artificial intelligence. *Economic Policy*, *40*(121), 225–256.
Krenz, A., Prettner, K., & Strulik, H. (2021). Robots, reshoring, and the lot of low-skilled workers. *European Economic Review*, *136*, 103744.
Kugler, A. D., Kugler, M., Ripani, L., & Rodrigo, R. (2020). *U.S. robots and their impacts in the tropics: Evidence from Colombian labor markets* (No. w28034). National Bureau of Economic Research.
Kuo, A., Manzano, D., & Gallego, A. (2024). Automation versus openness: Support for policies to address job threats. *Journal of Public Policy*, *44*(1), 1–23.
Kurer, T., & Gallego, A. (2019). Distributional consequences of technological change: Worker-level evidence. *Research & Politics*, *6*(1).

PROOFS

Leduc, S., & Liu, Z. (2024). Automation, bargaining power, and labor market fluctuations. *American Economic Journal: Macroeconomics*, *16*(4), 311–349.
Leighley, J. E., & Nagler, J. (2007). Unions, voter turnout, and class bias in the U.S. Electorate, 1964–2004. *Journal of Politics*, *69*(2), 430–441.
Lerch, B. (2021). *Robots and nonparticipation in the US: Where have all the workers gone?* Università della Svizzera italiana.
Liu, Y., Wang, H., & Yu, S. (2025). *Labor demand in the age of generative AI: Early evidence from the U.S. job posting data* (Policy Research Working Paper No. 11263). World Bank. http://hdl.handle.net/10986/44004.
Magistro, B., Borwein, S., Alvarez, R. M., Bonikowski, B., & Loewen, P. J. (2025). Attitudes toward artificial intelligence (AI) and globalization: Common microfoundations and political implications. *American Journal of Political Science*, *70*, 348–365.
Manning, S., Mishkin, P., Rock, D., & Eloundou, T. (2023). GPTs are GPTs: An early look at the labor market impact, potential of large language models. *arXiv Preprint arXiv*.
Margalit, Y., & Raviv, S. (2025). The politics of using AI in policy implementation: Evidence from a field experiment. *British Journal of Political Science*, *56*, e13.
McElheran, K., Yang, M.-J., Brynjolfsson, E., & Kroff, Z. (2025). *The rise of industrial AI in America: Microfoundations of the productivity J-curve(s)*. US Census Bureau, Center for Economic Studies.
Meyer, B. (2019). Financialization, technological change, and trade union decline. *Socio-Economic Review*, *17*(3), 477–502.
Microsoft. (2024). AI at work is here. Now comes the hard part.
Milanez, A. (2023). *The impact of AI on the workplace: Evidence from OECD case studies of AI implementation. OECD social, employment and migration working papers*. Paris: OECD.
Milner, H. V. (2021). Voting for populism in Europe: Globalization, technological change, and the extreme right. *Comparative Political Studies*, *54*(13), 2286–2320.
Milner, H. V. (2025). *AI's political disruptions: Political consequences of different winners and losers at varying rates of change*. Princeton University: Working Paper.
Mitts, T., & Ravis, S. (2025). *How media coverage and elite communication shape public opinion on AI regulation*. Tel Aviv University.
Narayanan, A., & Kapoor, S. (2025). *AI as normal technology*. Knight First Amendment Institute.
Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, *381*(6654), 187–192.
Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). The impact of AI on developer productivity: Evidence from GitHub Copilot. *arXiv preprint arXiv:2302.06590*.
Pinheiro, A., Sochirca, E., Afonso, O., & Neves, P. C. (2023). Automation and off(re) shoring: A meta-regression analysis. *International Journal of Production Economics*, *264*, 108980.
Pizzinelli, C., Panton, A. J., Tavares, M. M. M., Cazzaniga, M., & Li, L. (2023). *Labor market exposure to AI: Cross-country differences and distributional implications*. International Monetary Fund.
Pontusson, J., & Swenson, P. (1996). Labor markets, production strategies, and wage bargaining institutions: The Swedish employer offensive in comparative perspective. *Comparative Political Studies*, *29*(2), 223–250.

PROOFS

Potter, B. (2024, January 10). *What progress has there been in industrial robots? Construction physics*. www.construction-physics.com/p/what-progress-has-there-been-in-industrial

Prytkova, E., Petit, F., Li, D., Chaturvedi, S., & Ciarli, T. (2025). *The employment impact of emerging digital technologies*. NBER Chapters.

Rahman, K. S., & Thelen, K. (2019). The rise of the platform business model and the transformation of twenty-first-century capitalism. *Politics & Society*, *47*(2), 177–204.

Restrepo, P. (2025). We *won't be missed*: Work and *growth* in the AGI *world* (No. w34423). National Bureau of Economic Research.

Rodrik, D. (2018). New technologies, global value chains, and developing economies (No. w25164). National Bureau of Economic Research.

Rosenfeld, J. (2019). US labor studies in the twenty-first century: Understanding laborism without labor. *Annual Review of Sociology*, *45*, 449–465.

Sarto, A., Tabellini, M., & Faber, M. (2025). *Local shocks and internal migration: The disparate effects of robots and Chinese imports in the US*. Available at SSRN 5113046.

Stemmler, H. (2023). Automated deindustrialization: How global robotization affects emerging economies – Evidence from Brazil. *World Development*, *171*, 106349.

Susskind, R., & Susskind, D. (2022). *The future of the professions: How technology will transform the work of human experts*. Oxford University Press.

Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, *35*(6), 2769–2794.

Tan, J., & Thelen, K. (2025). Cloud capitalism and the AI transition. *Politics and Society*. 00323292251396395

Tomlinson, K., Jaffe, S., Wang, W., Counts, S., & Suri, S. (2025). Working with AI: Measuring the applicability of generative AI to occupations. *arXiv preprint arXiv*:2507.07935.

Webb, M. (2020). *The impact of artificial intelligence on the labor market*. Available at SSRN 3482150.

Yan, A. (2025). The minimal effects of union membership on political behavior. *Quarterly Journal of Political Science*, *20*(4), 537–568.

Yokley, E. (2025, December 1). Support for AI data center bans is growing. Morning Consult Pro. https://pro.morningconsult.com/analysis/ai-data-center-energy-prices-november-2025. Accessed on February 9, 2026.

Yu, Feiyang, Moehring, Alex, Banerjee, Oishi, Salz, Tobias, Agarwal, Nikhil, & Rajpurkar Pranav. (2024). Heterogeneity and predictors of the effects of AI assistance on radiologists. *Nature Medicine*, *30*(3), 837–849.

Zhang, B. (2022). No rage against the machines: Threat of automation does not change policy preferences. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 856–866). https://dl.acm.org/doi/pdf/10.1145/3514094.3534179.

PROOFS

8

AI, Geopolitics, and National Security

Sarah Kreps, Ben Buchanan, Michael Horowitz, and Erica Lonergan

PROOFS

8.1 INTRODUCTION

Artificial intelligence (AI) is reshaping international security. Unlike earlier military technologies, it does not introduce a single new weapon or platform. Instead, it changes how states generate, interpret, and exercise power. Earlier technological revolutions commonly studied in international relations transformed specific dimensions of warfare. Gunpowder expanded the reach of force. Nuclear weapons multiplied destructive potential. Cyber capabilities expanded the domain in which states could disrupt, surveil, and coerce. Artificial intelligence operates differently. It functions as a general-purpose technology that scales perception, prediction, and decision-making – activities historically constrained by human cognition and organizational capacity. By accelerating information processing, expanding analytical reach, and enabling new forms of automation across military and civilian domains, AI alters how power is produced and applied rather than simply adding new instruments of force.

This chapter argues that although autonomous weapons have attracted the most attention in debates about military AI, the broader geopolitical consequences of AI arise from the infrastructure, industrial systems, and institutional processes required to deploy AI at scale. Advantage increasingly depends on the ability to mobilize compute infrastructure, coordinate complex technological ecosystems, and integrate AI into organizational decision-making. In this sense, competition over AI resembles competition over industrial capacity and technological infrastructure rather than a traditional arms race centered on discrete weapon systems.

Disclosure: Generative AI tools were used during the drafting process for idea generation, organizational structuring, and minor language refinement. The authors retain full responsibility for the content, analysis, and interpretation presented in this work.

Tucker_9781047774734_C008.indd 247 08-Sep-26 2:07:36 PM

A fundamental challenge with AI is that it is dual use. Many of the most advanced AI systems are developed by private firms operating in commercial technology sectors, while military applications often depend on integrating those capabilities into existing command structures and operational practices. As a result, AI rivalry increasingly unfolds through control over semiconductor supply chains, compute infrastructure, data centers, and standards governing technological ecosystems. These dynamics are particularly visible in the rivalry between the United States and China, where competition extends beyond research laboratories and battlefield applications toward the material and industrial systems that enable AI development.

These characteristics also complicate the ability of states to assess relative technological capability, an important component of evaluating military power and strategic advantage. Performance depends simultaneously on algorithms, data, computational infrastructure, and organizational adoption. Some of the most visible systems operate in commercial settings, while many military applications remain classified. At the same time, AI capabilities spread rapidly through open-source ecosystems, commercial spillovers, and battlefield experimentation. Concentration and diffusion therefore occur simultaneously, producing persistent uncertainty about relative technological advantage.

Uncertainty, in turn, intensifies strategic competition. When states cannot easily assess rivals' technological progress, they often assume worst-case scenarios. Technological developments that appear experimental or defensive may be interpreted as signals of offensive capability. Under these conditions, AI amplifies familiar security dilemma dynamics even when underlying shifts in material power remain ambiguous. Governments are therefore competing not only over military applications but also over the industrial infrastructure and technological ecosystems that support AI development.

Understanding these developments requires situating AI within broader theories of technological change and international competition. Previous research shows that certain technologies transform global power distributions by imposing organizational and financial demands that only some states can meet (Posen 1984; Horowitz 2010). Artificial intelligence shares these characteristics but introduces an additional complication: Technological advantage depends not only on invention but also on institutional adoption, infrastructure coordination, and diffusion pathways that spread capabilities across a widening range of actors. The effects of AI therefore unfold unevenly across domains and states, producing layered patterns of concentration and diffusion that complicate traditional measures of military power.

This chapter examines how AI affects international security across four interrelated arenas. First, it explores how AI alters battlefield decision-making, military organization, and escalation dynamics through decision-support tools, autonomous systems, and doctrinal adaptation. Second, it analyzes how AI transforms great-power competition by shifting rivalry toward the industrial infrastructure and technological supply chains

PROOFS

Tucker_9781047774734_C008.indd 248 08-Sep-26 2:07:36 PM

that sustain frontier development, with a focus on the US–China relationship. Third, it examines the material foundations of AI capability, focusing on semiconductors, compute infrastructure, and the political economy of scaling advanced technological systems. Fourth, it analyzes how AI interacts with cyberspace and information warfare by transforming intelligence analysis, expanding opportunities for deception, and reshaping the offense–defense balance in digital conflict.

These dynamics also create significant governance challenges. Traditional arms-control frameworks rely on clear weapons categories, centralized production, and observable verification mechanisms. Artificial intelligence lacks these characteristics. Governance therefore emerges through indirect and fragmented approaches – including export controls, technical standards, voluntary industry coordination, and alliance-based cooperation – that attempt to shape technological trajectories rather than prohibit specific capabilities.

Finally, AI diffusion expands the range of actors capable of influencing international security. Commercial models, cloud infrastructure, and open-source software lower barriers to entry for military and intelligence innovation in many applications outside the frontier of model development. As a result, states increasingly compete not only with rival governments but also with technologically empowered private firms, insurgent organizations, and transnational cyber groups.

8.2 AI ON THE BATTLEFIELD AND IN GEOPOLITICS

8.2.1 AI and the Reconfiguration of Military Power

Technological change has long shaped the distribution of power in international politics. International relations scholarship has shown that certain innovations transform the underlying requirements for generating and projecting military force. Horowitz (2010) describes nuclear weapons and aircraft carriers as "system-altering" capabilities because they imposed steep organizational and financial barriers that only a limited number of states could overcome. Similarly, Buzan and Herring (1998) identify "transformational" military technologies as those that fundamentally restructure security competition by altering the cost, scale, or character of warfare.

Artificial intelligence increasingly appears to belong within this category, but for reasons that differ in important respects from earlier military revolutions. Unlike nuclear weapons, missiles, or conventional armaments, AI does not derive strategic significance primarily from destructive force. Instead, AI functions as a dual-use, general-purpose technology that operates across civilian and military sectors simultaneously. Advances in data analytics, machine learning, and automated decision-support systems improve intelligence processing, logistics optimization, targeting, cyber operations, and command and control rather than directly increasing explosive capability. In this sense, AI

PROOFS

reshapes military power less by replacing traditional weapons and more by transforming how states organize, integrate, and employ them.

This dual-use character complicates both strategic competition and governance. Because AI development is deeply embedded in commercial technology ecosystems, advances in civilian computing infrastructure, software engineering, and data management frequently translate into military capability. Scholars increasingly emphasize that AI's integration across civilian and defense sectors has the potential to alter patterns of geopolitical competition by generating new dependencies, creating novel vulnerabilities, and shifting the locus of military advantage toward states that can mobilize technological ecosystems rather than simply field superior weapon systems (Horowitz 2018; Vaynman and Volpe 2025). As a result, AI's strategic significance emerges not only in battlefield applications but also in global supply chains, export control regimes, alliance networks, and industrial policy debates.

The military and geopolitical effects of AI therefore operate through systemic and cumulative mechanisms rather than singular technological breakthroughs. Artificial intelligence may not transform warfare through a single decisive weapon comparable to nuclear arms, but its integration across sensing, analysis, and operational coordination has the potential to reshape how states generate and sustain military power over time. Understanding AI's strategic impact requires attention not only to autonomous systems or weapons platforms but to the broader technological, economic, and organizational ecosystems through which military capability is produced and exercised.

8.2.2 Horizontal Diffusion and Substitution Effects

In the context of national defense, AI use cases principally fall into three buckets: back-office administration, surveillance and decision support, and the battlefield. While back-office applications shape long-run military efficiency, the dynamics most salient for signaling, escalation, and competition emerge in decision support and battlefield integration, with the latter generating the greatest disruption to established military decision-making and raising unresolved questions about the scale and consequences of delegating lethal force.

Indeed, the back-office uses of AI for militaries look much as they do in other parts of government or the private sector – optimizing personnel processes, acquisition workflows, peacetime logistics, and force management. While these applications appear mundane, their strategic significance lies in their cumulative effects: Improved administrative efficiency reduces friction in force generation, accelerates mobilization, and allows militaries to sustain complex operations with smaller bureaucratic and labor footprints. Although such uses carry risks if systems are poorly trained or deployed, their symmetry with civilian applications makes them less distinctive from an international relations perspective (Global Commission on Responsible Artificial Intelligence in the Military Domain 2025).

PROOFS

A second area where militaries are integrating AI is decision support, computational tools that use machine learning and large-scale data analysis to generate predictions, recommendations, or risk assessments that inform rather than replace human decision-making in complex operational environments (US Department of War 2026). These systems typically aggregate inputs from multiple sensors and data streams, apply machine-learning models to filter, rank, and prioritize information, and present assessments or recommended courses of action to human operators within existing command-and-control structures. By reducing cognitive load and accelerating information processing, decision support tools reshape how commanders perceive and act on the battlefield.

The most prominent domain for such systems is intelligence, surveillance, and reconnaissance, where AI can process vast quantities of data more quickly than human analysts and identify patterns that would otherwise remain obscured. Decision-support tools also increasingly assist commanders in operational planning and, in some cases, targeting decisions (Nadibaidze, Bode, and Zhang 2024). These applications reconfigure military power by shifting advantage toward actors able to integrate data, analytics, and command structures effectively. Over time, this transformation may redistribute strategic advantage away from states that rely primarily on platform superiority or force size and toward those capable of sustaining complex data and algorithmic ecosystems.

These AI systems also compress decision timelines by accelerating information processing and operational planning cycles, potentially narrowing opportunities for deliberation during crises. Recent US military operations in Iran illustrate how this dynamic has appeared in practice. Anthropic's Claude AI was integrated into tools like the Maven Smart System to synthesize intelligence reporting, propose hundreds of potential targets, prioritize them by importance, and provide precise location coordinates, all within highly compressed planning timelines. This enabled over 1,000 strikes in the first twenty-four hours (Copp et al. 2026).

These systems raise concerns about automation bias – situations in which operators defer to algorithmic outputs beyond their demonstrated reliability. Experimental evidence suggests such bias may influence national security decision-making, particularly when users lack sufficient familiarity with the systems they employ (Horowitz and Kahn 2024). As reliance on AI-mediated analysis expands, such biases may introduce new pathways for misperception and escalation, especially in fast-moving conflict environments where verification and human oversight become more difficult to sustain.

The final category of military AI use lies on the battlefield itself in the context of lethal strike decisions. These dynamics are most visibly unfolding through drone warfare – understood here as the use of remotely piloted, semi-autonomous, and increasingly autonomous unmanned aerial systems for surveillance, targeting, and strike operations – which provides the clearest empirical illustration of how autonomy is being integrated into military operations.

PROOFS

Most deployed systems to date – including loitering munitions such as the Israeli Harop, which function as hybrids between drones and missiles that loiter while searching for targets before striking – delegate navigation or target recognition while preserving human decisions over lethal force.

Autonomous strike systems remain difficult to field when militaries seek reliable target identification under adversarial conditions while minimizing collateral damage and complying with international humanitarian law. Achieving this standard requires resilient operation in contested communications environments, real-time onboard computation within strict size and power constraints, and algorithmic decision-making capable of distinguishing lawful targets from civilians under conditions of uncertainty. However, these systems become considerably easier to deploy if the testing and evaluation shows these systems to be more effective than systems with more direct human supervision. The resulting trade-off highlights why autonomous strike capabilities remain both technologically and politically contested (Kaag and Kreps 2014).

These concerns are closely tied to ongoing debates about autonomous weapon systems. US Department of Defense guidance emphasizes that autonomous weapon systems must allow commanders and operators to exercise appropriate levels of human judgment over the use of force, while International Committee of the Red Cross discussions have emphasized the challenges autonomous systems pose for compliance with international humanitarian law including the principles of distinction, proportionality, and precautions in attack (U.S. Department of Defense 2023; ICRC 2014).

Yet, some worry that these legal and ethical commitments may prove difficult to sustain under conditions of military competition. If faced with a situation where an adversary deploys an autonomous system and perceives advantages of doing so, a state may feel pressure to reciprocate. Those advantages may be that they reduce the time of decision loops – from identifying to striking time-sensitive targets – or circumvent the need for real-time communication if electronic warfare has been disruptive. The increasing reliance, however, could create new risks. Adversaries can deploy decoys, spoof sensors, or deliberately mix civilian and military assets to manipulate automated targeting systems. Rapid machine-driven strikes can also accelerate escalation if autonomous systems misidentify targets or launch follow-on attacks without human reassessment. These dynamics illustrate how time pressure, survivability concerns, and competitive adaptation can push militaries toward lethal autonomy despite persistent uncertainty about legal compliance, escalation control, and accountability. How responsible militaries are at integrating AI will depend on operational demands and organizational politics as much as anything.

8.2.3 Diffusion, Assessment, and the Security Dilemma

Because military AI adoption unfolds in a competitive environment, states could interpret adversary investments in autonomy not only as technological

PROOFS

developments but also as potential shifts in relative military advantage. Even defensive or efficiency-driven deployments can generate uncertainty about adversaries' intentions and capabilities, particularly when autonomous systems reduce decision timelines or operate in communication-denied environments. Under such conditions, efforts to preserve deterrence credibility or avoid operational disadvantage can drive reciprocal adoption of AI-enabled systems, even when states recognize the risks of escalation or accidental conflict. These interactive dynamics shift the problem from one of technological capability alone to one of strategic competition under uncertainty.

The discussion raises the question of what it means to fall behind, which would be at the heart of understanding the consequences of AI for geopolitical competition and the balance of power. Horowitz (2018) argues that AI is better understood as a general-purpose technology – akin to electricity or the combustion engine – whose significance depends on how organizations adopt and apply it. Should capability be judged by algorithmic novelty, model scale, compute volume, or integration into command-and-control systems? The absence of clear thresholds complicates both competitive assessments and governance frameworks. Visible private-sector deployments like ChatGPT capture public attention, while public-sector classified systems remain obscured, widening the gap between perception and reality.

A world characterized by limited information and uncertainty about the trajectory of AI development – particularly whether advances could lead to Artificial General Intelligence (AGI) or superintelligence – creates incentives for worst-case thinking. If AI development includes a point at which recursive self-improvement produces rapid capability expansion, achieving that threshold first could plausibly confer significant strategic advantages. Even if such a discontinuity never materializes, the possibility that it might exist can influence state behavior. Importantly, the strategic consequences of AGI need not depend on precise technical definitions or even on its eventual realization; perceptions that such a capability threshold might exist can themselves drive competitive behavior. Because private firms, rather than governments, operate at the cutting edge of AI development, governments may struggle to forecast technological progress more than they have with traditional military systems. Under these conditions, states may interpret ambiguous AI advances by rivals as signals of potential offensive advantage, increasing the risk of security dilemmas.

As Robert Jervis (1978) argued in his seminal work on the security dilemma, states often respond less to what their rivals can do than to what they believe those rivals intend to do, and misperceptions in this domain can generate spirals of mistrust even in the absence of aggressive aims. Historical precedents illustrate how emerging technologies can intensify these dynamics. For example, the Reagan administration's Strategic Defense Initiative, though framed as a defensive effort to protect against nuclear attack, was widely interpreted

PROOFS

by Soviet leaders as a potential attempt to neutralize deterrence, contributing to intensified strategic competition despite significant uncertainty about the system's technical feasibility.

More recent work by Farrell and Newman (2019) extends this insight to an era of economic and technological interdependence, showing how control over networks, standards, and chokepoints can be interpreted simultaneously as defensive safeguards and as instruments of coercion. In this context, policies that are framed domestically as prudent risk management may be read abroad as hostile signals, reshaping expectations and accelerating rivalry even when underlying material balances change little.

The risk of hacking and spoofing algorithms, and the potential for algorithms to make decisions faster than humans can effectively review them, could further increase inadvertent escalation risks (Johnson 2021). Especially for states that fear the disruption or destruction of command-and-control leadership structures – even in nonnuclear contexts – the prospect of warfare at machine speed may encourage reciprocal adoption of AI-enabled decision making. Operational control by AI systems, particularly when algorithms are optimizing weapon selection and targeting, could generate escalation dynamics that become difficult to manage or reverse (Johnson 2022).

Alternatively, some critiques of the spiral model specifically suggest that concerns about AI-enabled escalation are overstated. Reiter (2025) argues that the desire of leaders for control over escalation pathways means that there are more offramps for war than the spiral model would suggest. Extending this logic to military AI, leaders may be reluctant to delegate consequential military operations or decisions to autonomous systems if doing so reduces their control over escalation (Reiter 2025).

8.3 THE US–CHINA AI COMPETITION

Nowhere is the interaction among technological uncertainty, competitive signaling, and concern about escalation control more salient than in the context of the US–China rivalry. It reflects both the uncertainties surrounding technological adoption and employment and the broader structural competition accompanying China's economic and military rise.

This section examines how competition over AI is shaped by material constraints, network position, and domestic political economy. It begins by contrasting the distinct visions through which the United States and China pursue AI leadership before turning to the narrow set of material inputs – on which success at the frontier depends. These inputs function as strategic chokepoints, generating leverage but not full control within global production networks. The analysis then considers how domestic political economy conditions each country's ability to mobilize and govern this infrastructure, linking these dynamics to the scaling logic of modern AI systems and to the semiconductor supply chains that anchor both the possibilities and limits of state influence.

PROOFS

8.3.1 Visions of AI Global Leadership

The United States and China signal the stakes of contemporary AI competition by treating leadership at the technological frontier as a strategic priority. In early 2026, China's President Xi Jinping met with his advisors and called AI "epoch-making technological transformation" that required a "whole-of-nation" approach to develop domestic capabilities to assert international leadership (Udinmwen 2026). The United States' current AI policy is called "Winning the Race: America's AI Action Plan" that suggests "Winning the AI race will usher in a new golden age of human flourishing, economic competitiveness, and national security for the American people" (White House 2025). The National Security Commission on Artificial Intelligence (2021) writes that "[w]e know that whoever translates AI developments into applications first will have the advantage. Now we must act."

While some AI competition is inevitable, some fear that framing AI development as a race toward dominance risks reinforcing competitive pressures. Security dilemma dynamics help explain why the rhetoric of an AI arms race persists despite the fragmented and multidimensional nature of AI development. When states portray AI leadership as a zero-sum contest, they encourage adversaries to interpret technological progress as evidence of offensive intent rather than defensive modernization. This perception can generate reciprocal investments in AI capability, even when advantage in one domain does not translate into overall technological dominance. In this sense, some fear that narratives of AI supremacy can operate as self-reinforcing signals that intensify rivalry, accelerate investment, and shift competition upstream toward control over the material and industrial foundations of AI development.

Complicating the AI competition debate is the possibility that no single state may achieve decisive technological dominance in AI. At present, some think that leadership is fragmenting across multiple domains, including frontier model development, control over computing infrastructure, influence over global technological standards, and the integration of AI into physical and military systems (Kahl 2026).

If AI competition unfolds across multiple domains rather than along a single technological dimension, the question becomes what conditions the prospects for success across those domains. Fragmented leadership does not eliminate structural constraints. It shifts attention to the material systems that enable innovation at scale. Understanding the trajectory of US–China AI competition therefore requires moving beyond rhetoric and institutional design to examine the infrastructural foundations that make frontier development possible.

8.3.2 Material Foundations of AI

Achieving leadership in AI depends on more than strategic intent or institutional design. It increasingly reflects control over the material and infrastructural foundations that underpin technological development. Insights gleaned from international political economy help explain why these foundations matter strategically.

PROOFS

As Farrell and Newman (2019) argue, power in globalized technological systems often arises from dominance over key nodes within production and governance networks rather than direct control over outcomes. Actors positioned at these hubs can shape the technological options available to others by structuring access to essential inputs, production processes, and infrastructure. Applied to AI, this framework directs attention to the components of AI development that exhibit high centrality, high switching costs, and concentrated governance. These characteristics increasingly define large-scale compute capacity, semiconductor manufacturing, and the infrastructure required to train and deploy advanced AI systems.

Among these inputs, compute has emerged as the binding constraint on AI development. Training and operating state-of-the-art models requires vast quantities of specialized hardware, dense networking capacity, reliable energy supply, and sustained capital investment. Unlike software or algorithmic innovations, which can diffuse relatively quickly, the physical infrastructure that underpins AI development depends on capital-intensive and slow-turnover production systems. Advanced chips must be designed, fabricated, and packaged through highly specialized manufacturing processes. Data centers must secure land, electricity generation, cooling systems, and network connectivity. Large-scale model training requires simultaneous access to all of these inputs over extended time horizons. As a result, improvements in AI capability increasingly track the ability to mobilize and coordinate physical infrastructure rather than isolated breakthroughs in model architecture.

The semiconductor industry sits at the core of this material foundation. Advanced semiconductor production has become concentrated among a small number of highly specialized firms embedded in a geographically dispersed supply chain, creating critical dependencies on a small number of firms and production centers (Miller 2022). Leading-edge chips are produced through manufacturing processes that only a small number of firms can execute at scale, using lithography and fabrication tools that themselves rely on highly complex and geographically concentrated supply chains. These characteristics make advanced semiconductors difficult to substitute and slow to reproduce, even with substantial financial investment.

Compute therefore functions not merely as an input to AI development but as a structural foundation that shapes the pace and direction of innovation. States and firms that command sustained access to advanced chips and the infrastructure required to deploy them gain the ability to train larger models, iterate more rapidly, and push performance boundaries outward. Those that lack such access face constraints on scale, training frequency, and deployment capacity, although advances in algorithmic efficiency and model optimization can partially mitigate these limitations.

Advances by Chinese AI developers, notably DeepSeek and Moonshot AI's Kimi models, illustrate both the potential and limits of cost-efficient innovation at the AI frontier. DeepSeek, for example, reportedly achieved competitive performance on several benchmark tasks using significantly lower training costs

PROOFS

Tucker_9781047774734_C008.indd 256 08-Sep-26 2:07:36 PM

and fewer computational resources than comparable Western frontier models, demonstrating that engineering optimization, data efficiency, and algorithmic refinement can partially offset raw compute scale. These developments suggest that high-end computing infrastructure is not the sole pathway to advancing model performance and that resource-constrained actors can narrow capability gaps through software and training innovations.

At the same time, some evidence from model evaluation and deployment patterns indicates that cost-efficient models often remain less reliable on complex reasoning tasks, less scalable across diverse enterprise and multimodal applications, and slower to iterate at frontier performance levels. High-compute models retain advantages in training speed, robustness, and the ability to support large-scale commercial deployment. Rather than eliminating the importance of compute, the experience of Chinese AI developers illustrate how AI competition increasingly involves a dynamic interaction between efficiency gains and infrastructure scale: Optimization can partially offset compute constraints, but sustained access to large compute clusters continues to define the outer limits of performance and iteration speed.

These material dependencies generate geopolitical significance not just because they are scarce, but because they occupy central positions within global production and governance networks. Control over compute infrastructure, semiconductor manufacturing, and data-center deployment functions as a chokepoint in the development of advanced AI capabilities. These inputs sit upstream of model development and deployment, function as bottleneck resources across national AI ecosystems, and remain governed by a small number of firms and jurisdictions. States that control access to these nodes can raise costs, slow diffusion, and redirect competitors' development pathways without directly suppressing innovation or relying on overt coercion. Conversely, states that depend on infrastructure governed by external actors face structural vulnerability even when they retain technical expertise and domestic research capacity.

Chokepoint power in AI operates gradually and indirectly. Restrictions on access to advanced chips or fabrication tools rarely halt technological progress outright. Instead, they reshape the conditions under which progress occurs by influencing the scale, speed, and direction of innovation over time. Competitors may pursue alternative hardware architectures, increase investment in efficiency-oriented research, or redirect development toward applications that require less computational intensity. In this sense, structural power in AI resembles other forms of economic statecraft in which influence arises through the structuring of technological opportunity rather than the prohibition of technological development itself.

Understanding AI competition therefore requires attention not only to who develops new models but also to who governs the infrastructure and production networks through which advanced AI becomes possible. Leadership in AI increasingly reflects the capacity to control, mobilize, and scale material

PROOFS

systems that shape technological trajectories across both military and commercial domains.

8.3.3 Scaling, Compute, and the Industrial Foundations of AI Power

The geopolitical significance of compute and semiconductor supply chains reflects a deeper technological dynamic within contemporary AI development. Progress at the frontier increasingly follows a scaling logic in which performance improvements derive from training larger models on greater quantities of data using expanded computational resources rather than from discrete conceptual breakthroughs. Under this logic, AI system performance improves as developers increase model size, training data, and compute allocation, even when underlying architectural innovations remain relatively modest. Countries such as Saudi Arabia and the United Arab Emirates are investing heavily in data centers and compute, using oil wealth to create a powerful AI hub in the Middle East (Satariano and Muzer 2025).

This pattern appeared in earlier machine-learning research but became especially visible with the emergence of large language models (LLMs). OpenAI's GPT-3, released in 2020, relied on design principles similar to earlier transformer architectures but expanded model parameters and training data by orders of magnitude. The resulting gains in language generation and reasoning ability demonstrated that expanding computational scale had become a primary driver of AI capability. Subsequent model development across the industry has largely reinforced this scaling trajectory.

Scaling dynamics generate strong incentives toward industrial concentration. Training advanced models requires access to specialized chips, massive data-center capacity, and sustained electricity supply, all of which exhibit high fixed costs and strong economies of scale. These requirements sharply limit the number of actors capable of competing at the frontier and reinforce the importance of firms' and states' abilities to mobilize large-scale capital investment and infrastructure deployment. As a result, AI development increasingly resembles other capital-intensive industrial sectors in which performance gains depend on the ability to expand and coordinate physical production systems rather than solely on intellectual or algorithmic innovation. Whether AI leadership will prove strategically decisive is unclear, but the belief that it might do so could drive states toward intensified technological rivalry.

The semiconductor ecosystem illustrates how scaling logic translates into industrial concentration. Leading AI chips depend on advanced fabrication processes that only a small number of firms can execute at scale. Taiwan Semiconductor Manufacturing Company (TSMC) occupies a uniquely central position in global semiconductor production as the dominant contract manufacturer of advanced logic chips, producing processors designed by leading AI

PROOFS

and computing firms and enabling the fabless production model that underpins much of the modern semiconductor industry.

Critical upstream suppliers such as Advanced Semiconductor Materials Lithography, the sole commercial supplier of extreme ultraviolet (EUV) lithography systems, provide the specialized equipment necessary for cutting-edge manufacturing. Extreme ultraviolet lithography enables the production of advanced transistors but requires extraordinary engineering precision and integrates supply chains spanning multiple countries, reinforcing barriers to entry in semiconductor production. The long development timelines and capital requirements associated with EUV systems create substantial barriers to entry, reinforcing concentration in semiconductor production and limiting the speed with which new competitors can emerge.

China's effort to build an indigenous semiconductor industry highlights the depth of the constraints embedded in advanced chip manufacturing. Beginning in the mid-2010s, Chinese leaders identified dependence on foreign semiconductors as a strategic vulnerability and launched a state-directed push to indigenize production through massive public investment, most notably via the "Big Fund." The "Big Fund," formally the China Integrated Circuit Industry Investment Fund, is China's primary state-backed financing mechanism designed to build a domestic semiconductor industry through large-scale public investment. As Miller (2022) documents, Chinese semiconductor initiatives have faced persistent difficulty replicating the precision manufacturing, supply-chain coordination, and tacit engineering expertise required for leading-edge chip production despite sustained funding and political backing. They remained dependent on foreign fabrication tools and equipment and, even when access remained available, they failed to achieve yields comparable to those of established producers. The experience highlights the long-time horizons and organizational complexity involved in semiconductor production, suggesting that manufacturing capability cannot be rapidly generated through capital investment alone.

Similar challenges appear even among advanced economies. Despite extensive incentives and close partnerships with TSMC, the United States has struggled to replicate the full manufacturing ecosystem and yield performance of TSMC's operations in Taiwan, including at facilities located on US soil (Goodman 2025). This pattern reinforces the extent to which advanced semiconductor production depends on deeply embedded industrial coordination, workforce specialization, and accumulated tacit knowledge.

Scaling dynamics do not produce absolute technological lock-in. Efficiency-oriented innovations can partially offset infrastructure constraints. Chinese AI firms, for example, have demonstrated the ability to produce competitive LLMs through cost-efficient training and algorithmic optimization strategies. Such developments suggest that constraints on compute often redirect innovation

PROOFS

toward alternative training approaches, cost reduction strategies, and model efficiency improvements rather than halting capability development altogether. However, scaling advantages remain significant because large compute clusters enable faster training iteration, broader multimodal integration, and more reliable deployment across complex commercial and military applications.

The scaling logic of AI therefore reinforces the strategic importance of compute infrastructure while simultaneously generating ongoing competition between scale and efficiency. This dynamic helps explain why AI development increasingly concentrates within a small number of firms and states while remaining open to partial diffusion through optimization and engineering innovation.

8.3.4 Semiconductors as Statecraft

The concentration of advanced semiconductor manufacturing and compute infrastructure has created new opportunities for states to treat technological supply chains as instruments of geopolitical strategy. Rather than functioning solely as commercial inputs, semiconductors increasingly operate as tools of economic statecraft through which governments attempt to shape the trajectory of technological development in rival states. In the context of AI, where access to large-scale compute capacity directly influences model training, deployment, and iteration speed, control over semiconductor supply chains has become a central lever of strategic competition.

By the late 2010s, the United States had begun to recognize advanced chip manufacturing as both a source of competitive advantage and a potential instrument of coercive leverage. China's repeated difficulties in indigenizing high-end semiconductor production – despite sustained public investment – suggested that restrictions on advanced chips and fabrication equipment could meaningfully slow China's progress in AI development. Early US policy measures under the first Trump administration targeted individual firms, reflecting uncertainty about the durability and scope of semiconductor leverage. Over time, however, policymakers increasingly concluded that control over the infrastructure supporting AI development, rather than individual technology transfers, would determine long-term competitive advantage.

This logic culminated in the Biden administration's October 2022 export controls, which restricted China's access to advanced AI chips and the equipment required to manufacture them. Subsequent coordination with allies expanded these restrictions to include lithography systems, semiconductor manufacturing tools, and advanced packaging technologies. Rather than functioning as traditional arms control, these measures sought to influence the pace, scale, and direction of Chinese technological development by limiting access to the physical infrastructure required to sustain AI training and deployment. In this sense, semiconductor export controls operate as a form of

PROOFS

indirect technological governance, shaping rival states' development pathways rather than prohibiting specific applications outright.

Yet the attempt to weaponize semiconductor chokepoints reveals important limits to structural technological power. Unlike nuclear or conventional military technologies, advanced AI hardware circulates through dense commercial ecosystems characterized by multinational supply chains, private-sector innovation, and dual-use applications. These characteristics make technological denial difficult to sustain and generate significant tradeoffs between security objectives and economic competitiveness. Restrictions on semiconductor exports can impose costs on domestic firms, incentivize supply-chain diversification by targeted states, and complicate alliance coordination when partner countries bear disproportionate economic burdens.

Policy debates within the United States reflect these tensions. The Biden administration's export control regime emphasized technological denial and long-term strategic constraint. By contrast, the Trump administration's subsequent decision to permit sales of high-performance Nvidia H200-class chips to Chinese customers reflected a more permissive approach that prioritized maintaining US commercial influence within global AI ecosystems. Advocates of partial export liberalization argue that allowing Chinese firms access to "good enough" computing infrastructure generates revenue for US companies, reinforces dependence on American software and programming platforms, and reduces incentives for China to accelerate indigenous semiconductor development. Critics counter that such policies risk narrowing the United States' compute advantage and accelerating Chinese AI capability development.

The broader experience of Chinese AI developers demonstrates that firms operating under partial infrastructure constraints can continue to produce highly competitive AI systems. While such models do not consistently match the performance of leading US frontier systems, they highlight how technological restrictions may redirect innovation toward efficiency, cost optimization, and alternative training strategies rather than preventing capability development altogether. These dynamics reinforce the reality that chokepoint power rarely produces absolute technological denial and instead operates by shaping competitive timelines and development pathways.

Semiconductor statecraft also carries broader geopolitical consequences because restrictions on technological access can deepen uncertainty about states' long-term intentions. A rival may interpret export controls as a temporary effort to slow technological convergence and preserve a short-term advantage or as evidence of a broader strategy to impose enduring constraints on its technological development. Uncertainty about which objective is driving policy can itself encourage states to reduce technological dependence and accelerate investment in domestic capabilities. Semiconductor restrictions may therefore reinforce the competitive dynamics they are intended to manage.

PROOFS

Finally, the effectiveness of semiconductor leverage depends on factors that extend beyond supply-chain control alone. Across both military and commercial technological revolutions, early pioneers have frequently struggled to translate technological breakthroughs into sustained institutional advantage. Leadership in AI may therefore depend less on exclusive access to hardware or models than on the capacity to absorb AI systems across military organizations, commercial industries, and public-sector institutions. As scholarship on general-purpose technologies suggests, long-term competitive advantage often accrues to actors that successfully integrate new technologies into economic production, governance systems, and operational practice rather than those that simply pioneer technological breakthroughs.

8.3.5 Domestic Political Economy and Infrastructure Mobilization

Although the United States and China confront many of the same material constraints at the AI frontier, their ability to manage and mobilize the underlying infrastructure differs sharply. These differences reflect broader features of domestic political economy and state capacity that condition how each country can pursue control over AI-related chokepoints.

China's strategy for developing AI involves coordinated policy and investment initiatives that extend beyond research firms to include the broader infrastructure supporting generative AI. In 2024, China's generative AI ecosystem was expanding rapidly, with more than fifty domestic developers of LLMs alongside major technology platforms and venture capital funding. They reflected a national emphasis on scaling both technological capabilities and the underlying compute infrastructure needed to support them (Triolo and Schaefer 2024). Unlike markets where infrastructure expansion is driven mainly by private hyperscalers, Chinese authorities have actively shaped the environment through regulatory approvals and institutional engagement that facilitate access to computing capacity, model deployment, and enterprise adoption. This state-aligned environment enables Chinese firms to iterate and deploy AI models while navigating constraints such as export controls on advanced chips. Although China's ecosystem remains diffuse compared with Western counterparts, the combination of public policy support, broad participation by domestic tech champions, and centralized regulatory oversight helps align investment and infrastructure development with national goals for technological competitiveness and economic integration.

The United States faces a different set of structural constraints in scaling AI infrastructure. Although US firms lead in model development, the physical expansion of data centers is shaped by decentralized regulatory and energy systems. Data-center siting is governed by local zoning processes, environmental reviews, water access constraints, and fragmented regional electricity markets that complicate large-scale infrastructure

PROOFS

planning. Empirical analyses of US data-center energy consumption indicate that rapid growth in compute demand is already placing increasing strain on regional power systems, requiring substantial upgrades to transmission capacity and generation resources to sustain continued expansion (Shehabi et al. 2024). More broadly, recent policy scholarship argues that the US model of privately driven AI innovation has reached structural limits, as infrastructure expansion, energy provisioning, and regulatory coordination increasingly require state involvement to sustain continued scaling (Buchanan and Collins 2025).

For much of the recent data-center expansion, local governments welcomed projects for their anticipated tax revenues and construction employment. As facilities have grown in scale and compute demand has increased, however, the local political calculus has begun to shift. Concerns over electricity costs, water usage, and grid reliability have generated greater scrutiny of whether communities absorb substantial infrastructure and environmental costs without commensurate long-term economic benefits. Sustaining further expansion may therefore require a development model that gives host communities a more direct stake in AI infrastructure through investments in local energy systems, workforce development, and community benefits (Kreps 2026a). Without such arrangements, growing local opposition may slow the infrastructure buildout on which continued AI scaling depends.

These emerging political headwinds are particularly consequential in the United States because federal authority over land use and electricity markets remains limited. Localized opposition can therefore introduce friction that national industrial or technology policy cannot easily override, creating the possibility that the United States could face bottlenecks in compute infrastructure deployment even while maintaining leadership in AI model innovation (Goetzel, Muro, and Methkupally 2026).

Those local debates have percolated into national-level political discussions, where elected leaders ranging from US Senator Bernie Sanders to Florida Governor Ron DeSantis, from different ends of the political spectrum, have called for either moratoria on AI data centers or expanded local authority to block their construction (Kimball 2026). Because data centers constitute the physical backbone of large-scale AI training and deployment, such domestic political conflicts can shape the scale and speed at which compute capacity expands. In turn, limits on compute expansion influence semiconductor demand, cloud infrastructure investment, and the broader technological ecosystems that underpin US–China strategic competition. Semiconductors have therefore shifted from a largely technical concern to a focal point of US–China strategic interaction that is also increasingly salient in domestic US politics. These debates illustrate how domestic political constraints may shape not only the pace of AI development, but the United States' capacity to sustain leadership in a globally competitive and infrastructure-intensive technological order.

PROOFS

8.3.6 Non-great Powers

Competition over AI capabilities is increasingly concentrated among great powers – particularly the United States and China – whose advantages rest on control over compute infrastructure, semiconductor supply chains, and large-scale model development. Yet dominance in these structural foundations does not translate into monopoly over military AI innovation. Many of the most visible battlefield applications of AI have emerged in conflicts involving middle powers and regional actors. This pattern reflects a broader dynamic in which military AI diffuses operationally across the international system even as frontier research and infrastructure remain concentrated among technologically advanced states.

This pattern of uneven diffusion occurs because military AI capabilities often depend less on frontier model development than on the integration of commercially available technologies, rapid doctrinal adaptation, and experimentation under combat conditions. While great powers dominate the upstream components of AI production, operational innovation frequently emerges among militaries actively engaged in conflict, where incentives to adopt emerging technologies, modify tactics, and accept operational risk are especially strong. Battlefield diffusion therefore demonstrates that structural technological leadership does not guarantee battlefield dominance, particularly in the early stages of technological transformation.

For advanced economies with relatively small populations – such as Israel, Singapore, and South Korea – AI-enabled systems help offset personnel constraints by automating or augmenting labor-intensive military functions. In areas such as intelligence analysis, surveillance, logistics coordination, and cyber operations, AI may allow states to increase the scale and tempo of operations without commensurate increases in personnel. Rather than competing at the technological frontier in model development, such states may derive operational advantage by rapidly integrating commercially available AI into existing military systems, adapting doctrine and training to exploit new capabilities, and reorganizing acquisition and command structures to accelerate deployment (Borchert et al. 2024). In this sense, AI diffusion is reshaping military effectiveness horizontally across the international system, reducing the premium on large national populations for certain military missions even as frontier advantages remain concentrated among major powers.

Israel's use of AI decision-support tools during the war in Gaza illustrates the operational appeal and controversy surrounding such systems. Designed to generate targeting recommendations at speeds beyond human analytic capacity, Israel's Lavender system reportedly provided significant operational efficiencies. Some observers argue that the system's speed and scale raised concerns about compliance with international humanitarian law, particularly when human operators lacked visibility into how targeting recommendations

PROOFS

were generated (Dorsey 2024). Other legal assessments emphasize that Lavender functioned as decision-support rather than autonomous targeting, with human operators retaining authority over strike decisions (Schmitt 2024). Regardless of these legal debates, the case highlights how AI systems can transform battlefield decision tempo even when human oversight remains formally intact.

The Russia–Ukraine war further demonstrates how states operating under resource and technological constraints can rapidly innovate in AI-enabled military systems (Chivers 2025). Ukraine has emphasized low-cost adaptation, integrating algorithmic tools for reconnaissance, targeting assistance, and drone coordination to offset conventional force disadvantages. Russia has pursued a more experimental approach, testing AI-enabled autonomy in unmanned ground vehicles and deploying algorithmic decision-support tools in cyber and electronic warfare operations, effectively using the conflict as a live laboratory for evaluating AI's battlefield utility (Konaev 2023). These cases illustrate how active conflict accelerates AI experimentation in ways that peacetime procurement and testing processes often constrain.

Despite this rapid operational innovation, most drone systems employed by both sides, consistent with the evidence above, remain remotely piloted or semi-autonomous rather than fully autonomous weapons. First-person-view drones function primarily as inexpensive remote sensing and strike platforms in which human operators retain real-time control. Their effectiveness derives less from advanced autonomy than from low cost, mass deployment, and persistent surveillance. During Operation Spider's Web, however, Ukrainian forces reportedly combined an open-source autopilot capable of autonomous operation when communications were disrupted with an AI-enabled targeting system trained to identify Russian bombers from three-dimensional scans (Horowitz, Kahn, and Schwartz 2025). The operation illustrates how autonomy is entering military systems incrementally, at the level of specific functions, rather than through an immediate transition to autonomous weapon systems.

This incremental shift may nevertheless be accelerating. At points in 2024 and 2025, battlefield assessments attributed an estimated 60–70 percent of frontline casualties to remotely piloted one-way attack drones (Santora et al. 2025; Chivers 2024). Continued electronic warfare and communications disruption have begun pushing both sides toward drones capable of greater onboard autonomy, suggesting a gradual shift toward more independent lethal systems when communication links degrade (Chivers 2025).

Battlefield diffusion also extends beyond active warfighting into surveillance and proxy warfare. India has incorporated AI into border monitoring along the Line of Actual Control with China, deploying automated image-recognition systems as part of its broader "AI in Defence" roadmap (Indian Ministry of Defence 2023; Bommakanti et al. 2024). Across the

PROOFS

Middle East, states such as the United Arab Emirates and Saudi Arabia have imported Chinese AI-enabled drone platforms, including the Wing Loong II and Rainbow CH-4, which have been deployed in proxy conflicts in Libya and Yemen (Kreps and Rogers 2025). These cases illustrate how AI capabilities can spread through arms exports, security partnerships, and commercial technology transfers, further accelerating operational diffusion beyond the states leading AI research.

These diffusion dynamics extend beyond state militaries. Non-state armed groups have demonstrated increasing ability to adapt commercially available AI-enabled tools, particularly in drone warfare, surveillance, and information operations. The relatively low cost of commercial drones, combined with open-source computer vision and navigation software, allows militant organizations and proxy forces to deploy capabilities that previously required state-level research and procurement infrastructure. Policy analysis suggests that rapidly improving and widely accessible AI models can generate "malicious uplift" by amplifying the scale and speed of harmful activities and enabling capabilities previously limited to advanced military or intelligence organizations to become available to smaller groups or even individuals. Such tools can accelerate operational planning, enhance targeting and surveillance capabilities, and expand information operations and cyber activities across a broader range of actors. While these groups rarely operate at the frontier of AI development, their ability to rapidly adopt and adapt commercially available technologies complicates deterrence, attribution, and escalation management by expanding the range of actors capable of deploying AI-enabled military capabilities (Chan et al. 2026).

Taken together, these cases demonstrate that military AI adoption follows multiple diffusion pathways, including organizational integration, wartime experimentation, export-driven proliferation, and surveillance-based deployment. This diffusion complicates efforts to assess relative military capability because states may achieve operational advantages through integration and experimentation even when they lack leadership in frontier model development or compute infrastructure. Battlefield diffusion therefore reinforces the broader challenge of measuring technological advantage in AI and underscores the growing disconnect between structural technological leadership and operational military effectiveness.

8.4 INFORMATION AND CYBER SUPERIORITY IN AN AI ERA

Artificial intelligence also intersects with international security through its effects on information power. As AI is arguably the most important "new" technology of the last generation, how cyber interacts with AI is therefore critical to understand. Artificial intelligence reshapes how states collect, process, and act on data, and it alters the conduct of cyber operations that depend on

PROOFS

digital infrastructure. This section examines two related but distinct questions: How does AI affect the balance of information power in intelligence and military decision-making? And how does AI influence the offense–defense balance in cyberspace?

In this chapter, cyberspace refers not merely to the public internet but also the broader digital ecosystem of networked information systems, cloud infrastructure, software platforms, data storage architectures, and communications networks. Artificial intelligence both depends on and transforms this environment. The same digital systems that enable AI-driven analysis also create new vulnerabilities to intrusion, manipulation, and deception.

8.4.1 AI and the Balance of Information Power

Artificial intelligence is transforming the balance of information power by expanding the speed, scale, and scope through which states collect, process, and interpret data. Since the end of the Cold War – particularly following the 1991 Persian Gulf War – information superiority has been viewed as a defining element of military and national power. Artificial intelligence intensifies this trend by enabling intelligence and military organizations to manage data volumes that exceed human analytic capacity.

Artificial intelligence–enabled analytic systems expand analysts' ability to process large volumes of unstructured data, automate translation and transcription, identify behavioral patterns, and detect emerging narratives across digital information environments (Horowitz and Lin-Greenberg 2022; Goldstein et al. 2023). These tools can also integrate information across bureaucratic stovepipes, allowing intelligence organizations to connect disparate data streams more rapidly and potentially reduce the risk of strategic surprise (Zegart 2007).

Artificial intelligence's significance in intelligence, therefore, lies less in revolutionary front-end breakthroughs than in its ability to restructure the analytic infrastructure that supports decision making. Artificial intelligence compresses analytic timelines, filters informational relevance, and standardizes interpretive steps that were previously manual, uneven, and labor intensive. These technologies do not replace human judgment, but they fundamentally condition it by shaping how information enters analytic pipelines, how quickly it moves through bureaucratic systems, and how it is implicitly prioritized.

Military organizations are increasingly integrating these capabilities into operational planning and battlefield decision making. The U.S. Department of Defense's Raven Sentry pilot program, for example, used AI trained on open-source data and historical insurgent attack patterns to predict likely insurgent activity in Afghanistan (Spahr 2024). More recent initiatives such as the Army Intelligence Data Platform seek to fuse intelligence from multiple sources and distribute it more rapidly to tactical commanders. Field experiments conducted

PROOFS

by the Army's 75th Innovation Command and the XVIII Airborne Corps, including exercises under the Maven Smart System and Scarlet Dragon programs, explore how AI tools can fuse sensor data, operate under degraded communication environments, reduce cognitive load on analysts, and support lethal and nonlethal targeting decisions.

Reflecting these developments, many policymakers and intelligence scholars argue that states capable of integrating AI across intelligence and operational decision-making processes may gain decisive informational advantages. Amy Zegart, for example, argues that AI offers potentially transformative capabilities by allowing organizations to process data at unprecedented scale, identify patterns more rapidly, and augment human analytic performance (Zegart 2022). Some observers have suggested that these technologies may even reduce or prevent intelligence failures by enabling faster detection of adversary intentions (Moran et al. 2023; but see Lonergan 2026).

8.4.2 The Vulnerability Paradox of AI-Enabled Information Power

Artificial intelligence–enabled information systems introduce new forms of fragility. Advanced models often function as opaque computational processes, making it difficult for analysts to interpret or audit machine-generated conclusions. They can also be brittle, performing reliably under familiar conditions while failing unpredictably when confronted with adversarial or novel inputs (Goldfarb and Lindsay 2021).

More significantly, AI systems create new attack surfaces within intelligence and military organizations. Adversaries may attempt to corrupt training datasets, manipulate model inputs, or exploit architectural weaknesses through cyber operations. Because AI systems increasingly operate within complex "systems of systems," malicious actors may target any component of the analytic pipeline. Machine-generated assessments may therefore become vulnerable to adversarial deception designed to manipulate decision-makers indirectly (Maschmeyer 2024).

These dynamics produce a paradox of AI-driven information power. As states rely more heavily on automated analytic systems to enhance situational awareness and decision speed, they may simultaneously increase exposure to deception strategies designed to exploit those systems. Informational advantage and informational vulnerability expand together.

8.4.3 AI and the Cyber Offense–Defense Balance

Artificial intelligence also reshapes competition in cyberspace itself. Here, the relevant question is whether AI shifts the offense–defense balance in digital conflict. Traditional offense–defense theory holds that when offensive capabilities are dominant, conflict becomes more likely because conquest appears easier (Jervis 1978; Van Evera 1998). This framework has long been debated,

PROOFS

particularly in cyber scholarship. Early analyses frequently described cyberspace as inherently offense-dominant due to speed, low barriers to entry, and ambiguity between espionage and attack (Buchanan 2016). More recent research challenges this view, showing that sophisticated offensive cyber campaigns require organizational capacity, skilled personnel, and sustained operational secrecy (Slayton 2016; Lonergan and Lonergan 2023; Smeets 2022). Cyber operations often succeed through deception and intelligence preparation rather than brute technical capability (Gartzke and Lindsay 2015).

Artificial intelligence may alter this balance in complex ways. Some analysts argue that AI strengthens cyber offense by accelerating attack planning, lowering technical barriers, and increasing the scale of operations. Machine learning can automate reconnaissance, generate adaptive malware, and speed exploitation processes that previously required manual intervention (Bonfanti 2022; Lohn 2025). A RAND analysis warned that AI-enabled cyber capabilities could enable a "splendid first cyber strike" capable of disrupting retaliatory responses (Mitre and Predd 2025).

At the same time, AI may enhance cyber defense. Artificial intelligence–enabled cybersecurity tools improve network monitoring, automate vulnerability detection and patching, and strengthen anomaly detection. Because defenders control network architecture, AI may allow them to dynamically reshape digital terrain and increase the cost of intrusion. Some scholars argue that AI may offer near-term advantages to defenders capable of integrating automated security systems effectively (Garfinkel and Dafoe 2019).

Empirical evidence suggests a mixed pattern. State-sponsored cyber actors increasingly use AI for reconnaissance and social engineering, yet human oversight remains central to strategic decision making. Artificial intelligence appears to increase automation and efficiency rather than eliminate the need for human control. As in the intelligence domain, AI reshapes competition incrementally rather than producing decisive shifts.

8.5 AI GOVERNANCE

8.5.1 Structural Constraints on AI Governance

Given its enormous implications for international politics, the spread of AI – especially frontier models and military applications – forces states to confront whether and how AI can be regulated, constrained, or stabilized through international mechanisms. Yet AI poses governance challenges that differ fundamentally from those associated with earlier arms control efforts. Unlike nuclear weapons, which depended on rare materials, centralized infrastructure, and observable tests, AI lacks chokepoints that reliably support verification and enforcement of binding international agreements. While AI development depends on upstream supply chains such as advanced semiconductors and

PROOFS

compute infrastructure that can serve as sites of strategic competition and export control, these chokepoints are deeply embedded in global commercial ecosystems and dual-use markets, making them far less suitable anchors for traditional arms control regimes.

Patterns of diffusion further complicate governance by shaping how states interpret one another's capabilities and intentions. Some AI models remain open-source and freely accessible, while others operate within closed, proprietary ecosystems that require paid access or state cooperation. Chinese developers have often released leading models that are open-source (Meinhardt et al. 2025). By contrast, many leading US frontier models remain proprietary, reflecting a model of concentrated control within private firms (Eastwood 2026). Open models accelerate diffusion, allowing states, firms, and non-state actors to reproduce capabilities quickly, but that same openness frustrates control, monitoring, and oversight. Closed systems, by contrast, concentrate advantage among a small number of firms and states and slow diffusion, yet they also restrict transparency and leave rivals uncertain about what those systems can do. Both pathways heighten mistrust in different ways: Openness raises fears of rapid proliferation, while closed development fuels suspicion of concealed advances.

As Vaynman and Volpe (2025) argue, these factors combine to create an "arms control dead zone," where the tools of past regimes – definitions, inspections, taboos, and treaties – have no clear anchor. Artificial intelligence governance thus begins from fragmentation rather than consensus, with states disagreeing not only about acceptable uses of AI but about where risks originate, whether from widespread access, concentrated control, or organizational misuse.

Not surprisingly, then, the governance landscape remains fragmented and has struggled to keep pace with the rate of technological change. Artificial intelligence lacks binding treaties, strong taboos, or shared verification mechanisms to date. Against this fragmented backdrop, governance initiatives rarely begin from shared premises. States disagree not only about acceptable uses of AI but also about where risks originate – whether from widespread access, concentrated control, or organizational misuse. As a result, governance initiatives rarely begin from shared premises. Instead of pursuing comprehensive regulation of AI as a technology, states and institutions advance narrower, partial approaches that reflect their own strategic priorities, regulatory traditions, and threat perceptions.

A further complication arises from the private-sector origins of frontier AI capabilities. Unlike earlier strategic technologies that governments helped develop through defense procurement, leading AI models have emerged primarily from commercial firms, leaving governments increasingly in the position of customers rather than sponsors (Kreps 2026b). As these systems become integrated into intelligence and military workflows, corporate decisions about model access, safeguards, and permissible uses can acquire national security

PROOFS

significance. Recent tensions between Anthropic and the Pentagon illustrate the resulting contest over authority: Private firms may seek to impose restrictions grounded in corporate safety policies, while governments resist allowing commercial providers to determine the permissible scope of military operations. The governance challenge therefore extends beyond coordination among states to the allocation of authority between governments and the firms that control increasingly strategic technological infrastructure.

8.7.1 Civilian/General AI Governance: Partial and Indirect Approaches

In response to these constraints, international AI governance has converged around a limited set of strategies, each attempting to impose structure on a technology that resists clear boundaries. One starting point has been scope and definition. Rather than regulate "AI" writ large, many proposals seek to delimit categories of particular concern. Some draw thresholds around model size or training compute, as in the European Union AI Act. It imposes stricter obligations on general-purpose models exceeding 10^{25} floating-point operations (FLOPs), a measure of the computational work used to train a model and a rough proxy for its scale and potential capabilities (EU AI Act 2024). By using training compute as a regulatory threshold, the Act seeks to identify models likely to pose systemic risks without relying on a fixed list of capabilities that may quickly become obsolete. The approach resembles the compute-based thresholds adopted in the Biden administration's 2023 executive order on AI, although that order was subsequently revoked by the Trump administration.

The difficulty is structural rather than technical. No widely accepted, objective metric currently captures "increased capability" across heterogeneous domains such as reasoning, autonomy, cyber exploitation, or biological knowledge. Static FLOPs cutoffs therefore function as administrable proxies, not definitive indicators of risk. As techniques evolve, fixed numerical thresholds risk becoming either overinclusive or obsolete, requiring continual revision rather than providing durable regulatory clarity.

A second cluster of proposals shifts attention from definitions to verification and transparency. Because training frontier systems requires substantial computational resources, some analysts have suggested monitoring the movement and use of advanced chips, analogizing compute to fissile material. Proposals include registries of large training runs, mandatory reporting requirements, and third-party audits of training processes or safety evaluations (Harack et al. 2023). In addition, some policy discussions have explored whether hardware-level mechanisms – such as embedded usage reporting, geofencing constraints, or remote management features built into advanced AI chips – could provide governments with greater visibility into how high-end compute is deployed. In principle, such mechanisms might allow exporting states to monitor or restrict the use of frontier chips abroad.

PROOFS

The goal of these approaches is not comprehensive enforcement but sufficient visibility to reduce mistrust and deter reckless behavior. Yet structural obstacles loom large. Unlike nuclear materials, advanced chips circulate through global commercial markets; firmware and system architectures can be modified; and firms may resist embedding intrusive monitoring features that threaten intellectual property or expose them to geopolitical retaliation. As RAND analysts observe (2023), nuclear verification succeeded because of identifiable chokepoints; AI supply chains, by contrast, are embedded in dense and highly distributed commercial ecosystems, making hardware-based transparency technically complex and politically fraught.

Recognizing these limits, a third set of proposals turns to new institutions and indirect levers. Some scholars and policymakers have advocated creating an international AI agency modeled on the International Atomic Energy Agency (IAEA) to monitor frontier models and mediate disputes (Belfield 2025). It is an idea echoed by the UK government at the 2023 Bletchley Park AI Safety Summit, through calls for a global AI Safety Institute network. Critics, however, argue that such an institution would struggle. As a general-purpose technology embedded in civilian markets, AI cannot be monitored or controlled in the same way as nuclear programs (Horowitz and Kahn 2025).

In practice, governance has increasingly shifted toward indirect and hybrid mechanisms. National policy instruments now function as de facto global governance tools. Export controls on advanced semiconductors – discussed earlier in the context of US–China competition – shape the pace and direction of AI development by constraining access to large-scale training infrastructure (Miller 2022). The Biden administration's October 2022 chip rules illustrate how licensing regimes tied to compute intensity can operate as forms of indirect control even without formal international agreement. Alongside these state-centered measures, governments have pursued multistakeholder approaches aimed at managing risk where binding regulation is infeasible.

In 2023, the White House secured voluntary commitments from leading AI firms on model evaluation, information sharing, and deployment practices. At the industry level, frontier developers themselves have moved to institutionalize coordination through the Frontier Model Forum, which brings together leading platform providers to develop shared safety practices, technical benchmarks, and approaches to risk assessment for the most advanced models.

Internationally, the Bletchley Park AI Safety Summit brought together governments and frontier-AI developers to coordinate research, share assessments of emerging risks, and develop common approaches to safety testing and evaluation. The UN High-Level Advisory Body on AI has similarly emphasized voluntary standards and coordination over legal compulsion.

Underpinning these initiatives is a growing reliance on norms and standards as substitutes for formal regulation. Technical benchmarks, safety evaluations, reporting practices, and certification schemes do not constrain capability development directly, but they shape expectations about responsible behavior and

PROOFS

acceptable risk. Over time, such standards can harden into de facto requirements through procurement rules, liability regimes, and market access, creating forms of governance that operate without treaty obligations or centralized enforcement.

Across these efforts, standards and norms increasingly serve as governance substitutes where treaties cannot. Trager et al. (2023) argue that licensing, certification, and oversight standards can function as a global scaffold for civilian AI governance by creating incentives for compliance through market access and reputational effects. Rather than restraining AI development directly, such mechanisms seek to shape behavior at key points of deployment and diffusion. Horowitz and Kahn (2025) argue that standards-setting approaches used for general-purpose technologies, such as those developed through the International Telecommunication Union, are more likely to succeed in the AI context than traditional arms control treaties.

The effectiveness of standards-based governance also depends heavily on sustained political commitment and international credibility. Periods of domestic policy volatility, shifting regulatory priorities, or skepticism toward multilateral coordination may complicate efforts to build durable global norms, while also creating space for competing governance models to emerge. As a result, standards-based approaches may represent a promising but politically contingent pathway for managing AI competition.

Taken together, these early efforts to apply international relations theories to the question of AI governance suggest that attempts at comprehensive regulation could founder on problems of definition, verification, and enforcement, as well as governance shifts toward leverage, coordination, and norm formation. This evolution reflects not a lack of ambition but an accommodation to the structural realities of AI as a general-purpose, dual-use technology embedded in global markets. Governance, where it occurs, is therefore partial, indirect, and contested – aimed at shaping trajectories and managing risk at the margins rather than controlling or banning, which have been out of reach. An interesting question for future researchers will be identifying the conditions in which these limits to governance approaches could change.

8.5.2.1 Military AI Governance: Application-Specific and Normative Efforts

If governance of civilian AI is fragmented and indirect, governance of military AI is even more constrained, though attempts at binding international legal regulation in specific areas are more advanced. Statutory regimes such as the EU AI Act largely sidestep military and national security uses, leaving governance efforts to concentrate on specific applications and norms rather than comprehensive control. Existing approaches to military AI governance fall into three broad categories. The first centers on use-case-specific regulation, most prominently the discussions on lethal autonomous weapon systems (LAWS) within the UN Convention on Certain Conventional Weapons (CCW). The LAWS

PROOFS

Group of Governmental Experts (GGE) represents an effort to isolate a particular application of AI and assess its compatibility with international humanitarian law (United Nations Office for Disarmament Affairs 2025). While these talks have clarified areas of concern, they have also revealed deep disagreement over definitions, scope, and the feasibility of meaningful restrictions, limiting progress beyond general principles. For example, the land mines and cluster munitions treaty efforts proceeded by agreement on a ban and then efforts to define specifically what was being banned. However, in a world where LAWS could encompass anything from loitering munitions to autonomous tanks, that approach has proven less successful as states have sought more clarity earlier in the process on what, specifically, would be regulated.

It is possible that these LAWS governance efforts could evolve in 2026, given the call by the UN Secretary General for binding regulations on LAWS use. Specifically, civil society groups are advocating for a binding treaty, whether through the United Nations or an ad hoc process such as the Treaty on the Prohibition of Nuclear Weapons. Consistent with international relations theories about civil society and transnational activism, if the CCW discussions continue to be slower than advocacy groups would like, civil society groups will likely attempt to rally supportive countries to convene an ad hoc process and negotiate a treaty to ban at least some uses of LAWS (Price 2003).

A second category encompasses broader efforts to articulate norms and expectations for military AI use without binding legal commitments. The Responsible AI in the Military Domain (REAIM) process exemplifies this approach. Through a series of summits, REAIM has brought together states, international organizations, and civil society to emphasize human responsibility, legal compliance, and risk awareness in military AI deployment. Related initiatives include the US-led Political Declaration on Responsible Military Use of AI and Autonomy, which has attracted more than fifty signatories, and the November 2024 US–China affirmation to maintain human control over nuclear weapons. These efforts do not impose constraints on capability development, but they seek to shape how AI is integrated into military decision making before practices become entrenched (US Department of State 2024; Renshaw and Hunnicutt 2024).

The third category focuses less on restraint and more on coordination among aligned states. Alliances and informal groupings such as NATO and the Quad have increasingly treated AI as an operational priority, emphasizing interoperability, shared standards, and collective learning. NATO has issued principles on responsible AI use, while the Quad's Critical and Emerging Technology Working Group has articulated shared commitments on AI development and deployment. These arrangements are not designed to limit AI use, but to ensure that allied militaries can operate together despite divergent national approaches (The Quad 2022; NATO 2021).

Across all three categories, a common constraint emerges. Unlike nuclear or chemical weapons, AI's general-purpose character and wide diffusion make

PROOFS

comprehensive, legally binding agreements hard to negotiate and even harder to enforce. A broad range of actors is adopting military AI, from major powers to middle states and non-state groups, each setting its own thresholds for acceptable use. As a result, states are experimenting in parallel rather than converging on binding shared rules. That said, they often are informally converging on shared norms for AI capability development, that is, the importance of testing and evaluation to ensure that capabilities work effectively. Even among allies, differences in confidence levels, testing standards, and targeting practices have already created interoperability challenges.

From an international relations perspective, this pattern suggests that traditional arms-control models offer limited guidance. Instead, confidence-building measures, transparency practices, and shared testing norms appear to be the most plausible building blocks for cooperation (Horowitz and Kahn 2025). Proposals include exchanges of information on military AI programs, joint exercises designed to probe failure modes, incident hotlines, and agreements governing interactions between deployed autonomous systems, modeled loosely on Cold War–era incidents-at-sea arrangements (CNAS 2023; GC REAIM 2025). These measures aim not to eliminate competition but to reduce the risks of accident, misinterpretation, and unintended escalation.

Efforts to govern military AI reflect a shift away from control toward management under conditions of diffusion and uncertainty. The challenge is no longer bilateral nor confined to weapons counts or deployment thresholds but one shaped by uneven adoption, organizational capacity, and opaque integration into command systems.

8.6 CONCLUSION

Artificial intelligence is reshaping international security not by introducing a discrete new instrument of force but by altering how power is generated, perceived, and exercised across economic, military, and informational domains. As a general-purpose technology embedded in civilian infrastructure, AI resists traditional categories of arms races and weapons competition. Its strategic significance lies less in measurable model performance than in how capabilities diffuse and are embedded across institutions. Across domains – from battlefield decision making to cyber operations and great-power rivalry – AI generates persistent ambiguity about relative advantage. States therefore compete and hedge even in the absence of clear shifts in material balance, amplifying familiar security-dilemma pressures without producing decisive dominance.

The analysis of military adoption highlights this point. Artificial intelligence–enabled systems are being incorporated across intelligence, decision support, surveillance, and semi-autonomous weapons, but in ways that are highly context dependent. States differ not only in access to models or compute but in organizational capacity, doctrine, and tolerance for risk. These differences shape how AI is actually used and, therefore, how much advantage it confers.

PROOFS

The chapter's discussion of battlefield integration and cyber operations suggests that AI's effects are mediated through institutions rather than determined by technology alone. In both domains, optimism about speed and automation is tempered by vulnerabilities to deception, brittleness, and organizational mismatch. The result is not a stable shift toward offense or defense, but a contingent balance shaped by how technologies are embedded in practice.

The examination of US–China competition highlights a parallel dynamic at the geopolitical level. Leadership in AI increasingly depends on access to compute and the infrastructure required to deploy it at scale, anchoring competition in semiconductor supply chains, energy systems, and industrial policy. These material foundations create chokepoints that offer leverage but not control. Export controls and allied coordination can slow and redirect development, but they operate indirectly and incompletely within global markets. At the same time, domestic political economy conditions shape each country's capacity to mobilize infrastructure, suggesting that long-run advantage will depend as much on governance and implementation as on innovation itself.

These features have important implications for AI governance. Traditional arms-control models, which rely on clear definitions, centralized production, and observable verification, offer limited guidance for a technology that is diffuse, rapidly evolving, and embedded in civilian systems. As a result, governance has shifted toward partial and indirect mechanisms: export controls, standards, voluntary commitments, and coordination among firms and allies. In the military domain, efforts have focused on application-specific regulation, norm articulation, and confidence-building measures rather than comprehensive prohibition. Across both civilian and military contexts, governance proceeds through fragmentation rather than consensus, reflecting disagreement not only about acceptable uses of AI but about where risks originate and how they should be managed.

Artificial intelligence competition is unlikely to yield clear equilibria or decisive technological victories. Instead, it will remain characterized by uneven adoption, contested signals, and ongoing uncertainty about capability, intent, and restraint. Advantage will accrue not simply to those who innovate first or control key inputs, but to those who can integrate AI systems effectively across institutions, manage their vulnerabilities, and interpret rivals' actions under conditions of limited information.

For scholars of international relations, this implies that AI should be understood less as a singular driver of change than as a force multiplier for existing structural and organizational dynamics. For policymakers, it suggests that efforts to govern AI will be incremental, contested, and shaped by trade-offs rather than comprehensive control. And for future research, it highlights the need to move beyond technological benchmarks toward systematic study of adoption, integration, and institutional learning – where the strategic consequences of AI are ultimately decided.

PROOFS

REFERENCES

Belfield, Haydn. 2025. "Domestic Frontier AI Regulation, an IAEA for AI, an NPT for AI, and a US-Led Allied Public-Private Partnership for AI: Four Institutions for Governing and Developing Frontier AI." arXiv preprint arXiv:2507.06379.

Bommakanti, Kartik, Yogesh Joshi, Shimona Mohan, Karthik Nachiappan, and Antara Vats. 2024. *AI in Defence and Border Monitoring: India's Use of Artificial Intelligence Along the Line of Actual Control*. Observer Research Foundation, May 10, 2024. www.orfonline.org/public/uploads/posts/pdf/20240510151332.pdf.

Bonfanti, Matteo E. 2022. "Artificial intelligence and the offence-defence balance in cyber security." *Cyber Security: Socio-Technological Uncertainty and Political Fragmentation*, edited by Myriam Dunn Cavelty and Andreas Wenger. London: Routledge: 64–79.

Borchert, Heiko, Torben Schütz, and Joseph Verbovszky, eds. 2024. *The Very Long Game: 25 Case Studies on the Global State of Defense AI*. Cham, Switzerland: Springer.

Buchanan, Ben. 2016. *The Cybersecurity Dilemma: Hacking, Trust, and Fear between Nations*. New York, NY: Oxford University Press.

Buchanan, Ben, and Tantum Collins. 2025. "The AI Grand Bargain: What America Needs to Win the Innovation Race." *Foreign Affairs*, November/December 2025. https://tinyurl.com/2s4vtv7n.

Buzan, Barry, and Eric Herring. 1998. *The Arms Dynamic in World Politics*. Boulder, CO: Lynne Rienner Publishers.

Chan, Kyle, Michael E. O'Hanlon, Qi Haotian, and Zheng Lefeng. 2026. *AI Risks from Non-State Actors*. Washington, DC: Brookings Institution.

Chivers, C. J. 2024. "How Suicide Drones Transformed the Front Lines in Ukraine." *The New York Times Magazine*, December 31, 2024, www.nytimes.com/2024/12/31/magazine/drones-weapons-ukraine-war.html.

Chivers, C. J. 2025. "In Ukraine, a New Arsenal of Killer A.I. Drones is Being Born." *The New York Times*, December 31, 2025, www.nytimes.com/2025/12/31/magazine/ukraine-ai-drones-war-russia.html.

CNAS (Center for a New American Security). 2023. *Artificial Intelligence and International Stability: Risks and Confidence-Building Measures*. Washington, DC: CNAS. https://tinyurl.com/by39aerw.

Copp, Tara, Elizabeth Dwoskin, and Ian Duncan. 2026. "Anthropic's AI tool Claude Central to U.S. Campaign in Iran, Amid a Bitter Feud." *The Washington Post*, March 4, 2026, www.washingtonpost.com/technology/2026/03/04/anthropic-ai-iran-campaign.

Dorsey, Jessica. 2024. "Israel's AI-Enabled Targeting of Hamas Members Jeopardizes Moral and Legal Standards of Warfare." *Utrecht University*, July 18, 2024. https://tinyurl.com/wdf53afa.

Eastwood, Brian. 2026. "AI Open Models Have Benefits. So Why Aren't They More Widely Used?" *MIT Sloan Management Review*, January 20. https://tinyurl.com/4r5kukca.

European Union. 2024. *Regulation (EU) 2024/1689 of the European Parliament and of the Council on Artificial Intelligence (AI Act)*. Official Journal of the European Union, L. 275.

PROOFS

Farrell, Henry, and Abraham L. Newman. 2019. "Weaponized Interdependence: How Global Economic Networks Shape State Coercion." *International Security* 44(1) (Summer): 42–79.

Garfinkel, Ben, and Allan Dafoe. 2019. "Artificial Intelligence, Foresight, and the Offense-Defense Balance." *War on the Rocks*. https://tinyurl.com/ycy2rxk3.

Gartzke, Erik, and Jon R. Lindsay. 2015. "Weaving Tangled Webs: Offense, Defense, and Deception in Cyberspace." *Security Studies* 24(2): 316–348.

Global Commission on Responsible Artificial Intelligence in the Military Domain. 2025a. "Global Commission on Responsible Artificial Intelligence in the Military Domain (GC REAIM)." *The Hague*, Netherlands: Hague Centre for Strategic Studies.

Global Commission on Responsible Artificial Intelligence in the Military Domain. 2025b. "Responsible by Design: Strategic Guidance Report on the Risks, Opportunities, and Governance of Artificial Intelligence in the Military Domain." *The Hague*, Netherlands: Global Commission on Responsible Artificial Intelligence in the Military Domain.

Goetzel, D., M. Muro, and S. Methkupally 2026. *Turning the Data Center Boom into Long-term, Local Prosperity*. Washington, DC: Brookings Institution.

Goldfarb, Avi, and Jon R. Lindsay. 2021. "Prediction and Judgment: Why Artificial Intelligence Increases the Importance of Humans in War." *International Security* 46(3): 7–50.

Goldstein, J. A., G. Sastry, M. Musser, R. DiResta, M. Gentzel, and K. Sedova 2023. "Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations." *arXiv*. https://doi.org/10.48550/arXiv.2301.04246.

Goodman, Peter. 2025. "18,000 Reasons It's So Hard to Build a Chip Factory in America." *New York Times*, www.nytimes.com/2025/12/04/business/tsmc-phoenix-fab.html.

Harack, Ben, Robert F. Trager, Anka Reuel, David Manheim, Miles Brundage, Onni Aarne, Aaron Scher, Yanliang Pan, Jenny Xiao, Kristy Loke, Sumaya Nur Adan, Guillem Bas, Nicholas A. Caputo, Julia C. Morse, Janvi Ahuja, Isabella Duan, Janet Egan, Ben Bucknall, Brianna Rosen, Renan Araujo, Vincent Boulanin, Ranjit Lall, Fazl Barez, Sanaa Alvira, Corin Katzke, Ahmad Atamli, and Amro Awad. 2025. *Verification for International AI Governance*. Oxford Martin AI Governance Initiative, 3 July 2025.

Horowitz, Michael C. 2010. *The Diffusion of Military Power: Causes and Consequences for International Politics*. Princeton, NJ: Princeton University Press.

Horowitz, Michael C. 2018. "Artificial Intelligence, International Competition, and the Balance of Power." *Texas National Security Review* 1(3): 36–57. https://tinyurl.com/4zmfceun.

Horowitz, Michael C., and Erik Lin-Greenberg. 2022. "Algorithms and Influence Artificial Intelligence and Crisis Decision-making." *International Studies Quarterly* 66(4): sqac069.

Horowitz, Michael C., and Lauren A. Kahn. 2024. "Bending the Automation Bias Curve: A Study of Human and AI-Based Decision Making in National Security Contexts." *International Studies Quarterly* 68(2): sqae020. https://doi.org/10.1093/isq/sqae020.

Horowitz, Michael C., and Lauren A. Kahn. 2025. "Nuclear Non-Proliferation Is the Wrong Framework for AI Governance." *AI Frontiers*, June 27.

PROOFS

Horowitz, Michael C., Lauren A. Kahn and Joshua A. Schwartz. 2025. "What Drones Can—and Cannot—Do on the Battlefield: The Pentagon Should Learn from Israel and Ukraine." *Foreign Affairs*, 4 July 2025, www.foreignaffairs.com/united-states/what-drones-can-and-cannot-do-battlefield.

Indian Ministry of Defence 2023. *AI in Defence: Roadmap for Adoption of Artificial Intelligence in the Indian Armed Forces*. New Delhi: Ministry of Defence. www.ddpmod.gov.in/sites/default/files/2023-11/ai.pdf

International Committee of the Red Cross. 2014. *Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects*. Geneva, Switzerland: International Committee of the Red Cross.

Jervis, Robert. 1978. "Cooperation Under the Security Dilemma." *World Politics* 30(2): 167–214. https://doi.org/10.2307/2009958.

Johnson, James. 2021. "'Catalytic Nuclear War' in the Age of Artificial Intelligence & Autonomy: Emerging Military Technology and Escalation Risk between Nuclear-armed States." *Journal of Strategic Studies* 45(4): 1–41.

Johnson, James. 2022. "Inadvertent Escalation in the Age of Intelligence Machines: A New Model for Nuclear Risk in the Digital Age." *European Journal of International Security* 7(3): 337–359.

Kaag, John, and Sarah Kreps. 2014. *Drone Warfare*. New York, NY: Polity Press.

Kahl, Colin H. 2026. "The Myth of the AI Race: Neither America Nor China Can Achieve True Tech Dominance." *Foreign Affairs*, January 12, 2026. www.foreignaffairs.com/united-states/myth-ai-race.

Kimball, Spencer. 2026. "Bernie Sanders and Ron DeSantis Speak Out Against AI Data Centers Over Electricity Prices." *CNBC*, January 1, 2026, https://tinyurl.com/bdf2jv2h.

Konaev, Margarita. 2023. "Use of AI and Autonomous Technologies in the War in Ukraine." *CNA*, October 2023, https://tinyurl.com/2vazsdt9.

Kreps, Sarah. 2026a. "A Better Way to Build AI." *Foreign Affairs*, July 6. www.foreignaffairs.com/united-states/better-way-build-ai.

Kreps, Sarah. 2026b. "The Integrated Circuit and the Future of American AI Leadership." *War on the Rocks*, July 3. https://tinyurl.com/3e6vy773.

Kreps, Sarah, and James Patton Rogers. 2025. *Drones: What Everyone Needs to Know*. Oxford, UK: Oxford University Press.

Lohn, Andrew J. 2025. "The Impact of AI on the Cyber Offense-Defense Balance and the Character of Cyber Conflict." *arXiv*. https://arxiv.org/abs/2504.13371.

Erica D. Lonergan, "The paradox of intelligence failures in the age of artificial intelligence," *Intelligence and National Security* (2026): 1–30.

Lonergan, Erica D., and Shawn W. Lonergan. 2023. *Escalation Dynamics in Cyberspace*. New York, NY: Oxford University Press.

Maschmeyer, Lennart. 2024. *Subversion: From Covert Operations to Cyber Conflict*. New York, NY: Oxford University Press.

Meinhardt, Caroline, Sabina Nong, Graham Webster, Tatsunori Hashimoto, and Christopher Manning. 2025. *Beyond DeepSeek: China's Diverse Open-Weight AI Ecosystem and Its Policy Implications*. Stanford, CA: Stanford Institute for Human-Centered Artificial Intelligence.

Miller, Chris. 2022. *Chip War: The Fight for the World's Most Critical Technology*. New York, NY: Scribner.

Mitre, Jim, and Joel B. Predd. 2025. "Artificial General Intelligence's Five Hard National Security Problems." *RAND Corporation*. https://tinyurl.com/5bawb7mn.

PROOFS

Moran, Christopher R, Joe Burton, and George Christou. 2023. "The US Intelligence Community, Global Security, and AI: From Secret Intelligence to Smart Spying." *Journal of Global Security Studies*, 8(2): ogad005.

Nadibaidze, Anna, Ingvild Bode, and Qiaochu Zhang. 2024. *AI in Military Decision Support Systems: A Review of Developments and Debates*. Odense, Denmark: Center for War Studies, University of Southern Denmark.

National Security Commission on Artificial Intelligence. 2021. *Final Report*. Washington, DC: National Security Commission on Artificial Intelligence.

NATO. 2021. "NATO Secretary General: Artificial Intelligence Will Be Key to Alliance Security." *NATO Press Release*, October 2021. www.nato.int/cps/en/natohq/news_187607.htm.

Posen, Barry R. 1984. *The Sources of Military Doctrine: France, Britain, and Germany Between the World Wars*. Ithaca, NY: Cornell University Press.

Price, Richard. 2003. "Transnational Civil Society and Advocacy in World Politics." *World Politics* 55 (4): 579–606.

RAND Corporation. 2023. *Insights from Nuclear History for International AI Governance*. Santa Monica, CA: RAND. www.rand.org/pubs/perspectives/PEA3652-1.html.

Reiter, Dan. 2025. *Untied Hands: How States Avoid the Wrong Wars*. Cambridge, UK: Cambridge University Press.

Renshaw, Jarrett, and Trevor Hunnicutt. 2024. "Biden, Xi Agree That Humans, Not AI, Should Control Nuclear Arms." *Reuters*, November 16.

Santora, Marc, et al. 2025. "A Thousand Snipers in the Sky: The New War in Ukraine." The *New York Times*, March 3, 2025, www.nytimes.com/interactive/2025/03/03/world/europe/ukraine-russia-war-drones-deaths.html.

Satariano, Adam, and Paul Mozur. 2025. "Saudi Arabia's New Power Play is Exporting A.I. to the World." *The New York Times*, November 4, 2025. www.nytimes.com/2025/10/27/technology/saudi-arabia-ai-exporter.html.

Schmitt, Michael N. 2024. "Israel – Hamas 2024 Symposium – The Gospel, Lavender, and the Law of Armed Conflict." *Articles of War, Lieber Institute for Law and Warfare*, June 28, 2024. https://lieber.westpoint.edu/gospel-lavender-law-armed-conflict/.

Shehabi, Arman, et al. 2024. *United States Data Center Energy Usage Report*. Berkeley, CA: Lawrence Berkeley National Laboratory.

Slayton, Rebecca. 2016. "What is the Cyber Offense-defense Balance? Conceptions, Causes, and Assessment." *International Security* 41(3): 72–109.

Smeets, Max. 2022. *No Shortcuts: Why States Struggle to Develop a Military Cyber-Force*. New York, NY: Oxford University Press.

Spahr, Thomas W. 2024. "Raven Sentry: Employing AI for Indications and Warnings in Afghanistan." *The US Army War College Quarterly: Parameters* 54(2): 95–109.

Trager, Robert, Ben Harack, Anka Reuel, Allison Carnegie, Lennart Heim, Lewis Ho, Sarah Kreps, Ranjit Lall, Owen Larter, Seán Ó. hÉigeartaigh, Simon Staffell, José Jaime Villalobos, et al. 2023. *International Governance of Civilian AI: A Jurisdictional Certification Approach*. Oxford, UK: Oxford Martin AI Governance Initiative and Centre for the Governance of AI.

Triolo, Paul, and Kendra Schaefer. 2024. "China's Generative AI Ecosystem in 2024: Rising Investment and Expectations." *National Bureau of Asian Research*, June 27, 2024. https://tinyurl.com/4d6hmt2c.

PROOFS

The Quad. 2022. "Quad Principles on Technology Design, Development, Governance, and Use." *Critical and Emerging Technology Working Group*, May 2022. https://tinyurl.com/rfffshe3.

Udinmwen, Efosa. 2026. "Xi Jinping Pushes China's AI Ambition But Warns Against Idle Capacity." *Tech Radar*, January 30, 2026. www.yahoo.com/news/articles/xi-jinping-pushes-chinas-ai-223500014.html.

UK Government. 2023. "Bletchley Declaration on AI Safety." November 1, 2023. www.gov.uk/government/publications/bletchley-declaration.

U.S. Department of Defense. 2023. *Autonomy in Weapon Systems. Department of Defense Directive 3000.09*. Washington, DC: U.S. Department of Defense.

U.S. Department of State. 2024. "Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy." Bureau of Arms Control, Deterrence, and Stability.

U.S. Department of War. 2026. "Artificial Intelligence Strategy for the Department of War." *Memorandum*, January 9, 2026. Artificial Intelligence Strategy for the Department of War. https://tinyurl.com/3dzvd4u3.

United Nations Office for Disarmament Affairs. 2025. "Group of Governmental Experts on Lethal Autonomous Weapons Systems." In *Convention on Certain Conventional Weapons*. Geneva, Switzerland: United Nations Office for Disarmament Affairs.

Van Evera, Stephen. 1998. "Offense, Defense, and the Causes of War." *International Security* 22(4): 5–43.

Vaynman, Jane, and Tristan Volpe. 2025. "Competition and Collusion: How the AI Arms Race Can Motivate Governance." In *RAND–Perry World House Collaboration Report*. Santa Monica, CA: RAND Corporation and Perry World House.

White House. 2025. "Winning the Race: America's AI Action Plan. The White House, Executive Office of the President." www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf.

Zegart, Amy. 2007. *Spying Blind: The CIA, the FBI, and the Origins of 9/11*. Princeton, NJ: Princeton University Press.

Zegart, Amy. 2022. *Spies, Lies and Algorithms: The History and Future of American Intelligence*. Princeton, NJ: Princeton University Press.

Proofs

9

AI and Political Theory

Linda Eggert, Ting-an Lin, Jeffrey Howard,
Lorenzo Manuali, and Rob Reich

PROOFS

9.1 INTRODUCTION

9.1.1 AI and Political Theory

Political theory is a rich and varied discipline, extending across philosophy, political science, international relations, law, and economics. Our focus in this chapter is on political theory as a predominantly normative – rather than empirical or historical – enterprise, concerned with assessing and defending values, principles, and reasons for action (Estlund, 2012; Miller, 2022).

The aim of this chapter is not to defend substantive positions but to set out some of the issues related to artificial intelligence (AI) that preoccupy political theorists. The task is to map emerging scholarship, identifying some of the most significant contributions to date, and to highlight propitious lines of inquiry for future research. In some cases, the task involves assessing what existing theories imply for new policy questions raised by AI. In other cases, it involves rethinking basic theoretical commitments or assumptions. What makes AI a fertile and important topic for political theorists is its potential to contribute to, or prompt us to rethink, our understanding of substantive problems that political theory has traditionally sought to address.

The chapter is divided into three main sections: (Section 9.2) "Justice," (Section 9.3) "Democracy," and (Section 9.4) "Rights and the Future." It does not discuss many other important issues of interest to political theorists, some of which are taken up in other chapters in this volume. These include issues concerning national security and international relations (Chapter 8), economic inequality and the labor force (Chapter 7), policymaking and the administrative state (Chapter 4), and research methods (Chapter 10).

Within the confines of this chapter, we cannot hope to adequately characterize the rich array of different subdisciplinary traditions and approaches

Tucker_9781047774734_C009.indd 282 08-Sep-26 2:11:04 PM

in political thought and theory. But this may be just as well. AI's claim on the attention of political theorists of all stripes has begun to blur traditional boundaries between different methods and approaches (Risse, 2023). While thinkers outside of analytic liberal-egalitarian political theory have long been concerned with the ways in which technology affects people's lives, much of contemporary analytic political theory is comparatively late to the party (Bernholz et al., 2021; Risse, 2023; Lazar, 2025). As political theory evolves as a field, AI may become a powerful catalyst for the convergence of different approaches within political thought.

9.2 JUSTICE

9.2.1 Algorithmic Injustice

This section provides an overview of scholarly discussions of *algorithmic injustice*, broadly construed as being concerned with situations in which AI systems introduce new forms of injustice or exacerbate existing ones.

Regarding the unjust impact of algorithms, considerable attention has been drawn to the issue of *algorithmic bias*: Algorithms produce biased outcomes that perpetuate or exacerbate existing forms of discrimination and oppression (O'Neil, 2017; Noble, 2018; Benjamin, 2019; Buolamwini & Gebru, 2018). Algorithmic biases occur, for example, when algorithms systemically assign higher health needs to white populations compared to people of color (Obermeyer et al., 2019) or when algorithms associate denigrating messages with women and people of color (Noble, 2018).

In response to concerns about algorithmic bias, the term *algorithmic fairness* was introduced to describe the vision of ensuring that algorithms operate fairly. (For an overview, see Hellman, 2025.) Initial scholarly debates differentiated various notions of fairness – primarily understood as ensuring statistical parity – and debated their suitability in assessing whether algorithms operate fairly (Kleinberg et al., 2016; Narayanan, 2018; Binns, 2020; Hedden, 2021; Hellman, 2025). Valuable critiques have been raised about the limitations of understanding algorithmic fairness as ensuring statistical parity. One prominent worry is that such an analysis problematically isolates AI systems from broader social contexts and overemphasizes technocentric responses (Hoffmann, 2019; Le Bui & Noble, 2020). These reflections have led to a "structural turn" that goes beyond exploring technical methods for debiasing AI systems and considers how AI systems shape broader power dynamics and contribute to injustices (Kalluri, 2020; Le Bui & Noble, 2020).

Major questions concerning algorithmic fairness include: What is algorithmic fairness? Why is it valuable? What does it require in practice? Different traditions of political theory have been drawn on to provide insights.

John Rawls's theory of justice (1971) and its implications for distributive justice have been widely consulted and translated into practical guidelines by

PROOFS

computer scientists. In political theory, Gabriel (2022) is a notable example of engagement with Rawlsian principles of justice that outlines their implications for AI. For example, what is known as Rawls's first principle of justice requires that the design and deployment of AI should protect the equal basic liberties of all. The principle of fair equality of opportunity implies moving away from a formal conception of fairness (or "de-biasing" AI systems) toward more substantive fair outcomes, such as how AI tools might help mitigate discriminatory impacts due to social membership. Furthermore, Gabriel suggests that the difference principle implies that, when AI systems are integrated into key social practices, they must work to the greatest advantage of the least advantaged members of society.

Scholarship in feminist and critical race theories offers a more multifaceted understanding of algorithmic (in)justice that extends beyond distributive justice. While the terminology varies, a common theme in this literature is its emphasis on power dynamics, social relations, and social status across different populations. This emphasis has led to suggestions to rethink the goals and practices of algorithmic fairness. For example, Iris M. Young's influential work on structural injustice (2010) has led scholars to reconceptualize the goal of algorithmic fairness as the pursuit of a more just social-technical structure in which unjustified power imbalances are mitigated (Lin and Chen, 2022; Kasirzadeh, 2022). Insights from intersectionality research (Crenshaw, 1991; Collins, 2002) and relational egalitarianism (Anderson, 1999; Fraser, 1995) suggest that the goal of algorithmic fairness should encompass not only distributive equality but also people's ability to relate to one another as equals (Kong, 2026).

Concerning practices of pursuing algorithmic fairness, reflections on power dynamics have led to advocacy for participatory approaches to technological development (Costanza-Chock, 2018; D'Ignazio & Klein, 2020), emphasizing the importance of engaging a wide range of stakeholders (especially with marginalized perspectives) in the process. Though they offer a promising direction, participatory approaches in AI design often face practical challenges and lack concrete guidelines. They have also raised concerns about "participation washing" (Birhane et al., 2022). How to best implement participatory visions remains an important question for future exploration.

Another body of literature concerns how AI models shape epistemic norms and practices and produce unjust impacts. One major theme is how AI systems contribute to *epistemic injustice* – that is, injustice concerning the domain of knowledge. Studies have suggested that the algorithms behind social media exacerbate epistemic injustice through algorithmic targeting and sorting (Stewart et al., 2022) and constitute a novel form of hermeneutical injustice, as algorithmic profiling leads to epistemically fragmented environments (Milano & Prunkl, 2025).

The content-creation capacity of generative AI models has further impacted collective knowledge and epistemic practices, presenting new forms of testimonial and hermeneutical injustices through amplification of misinformation,

PROOFS

Tucker_9781047774734_C009.indd 284 08-Sep-26 2:11:04 PM

intentional manipulation of wrongful content, and increased hermeneutical obstacles for marginalized groups (Kay et al., 2025; Mollema, 2025). Since algorithmic systems tend to encode and manifest the conventions of dominant groups, such as accents and communication styles, the constant cultural code-switching required of marginalized groups constitutes a new form of epistemic oppression: cultural smothering (Falbo & LaCroix, 2022). Similar concerns have been raised about AI models' tendency to generate content that reflects or prioritizes Western-centric perspectives, which may unjustly reinforce Western epistemic hegemony (see the Section 9.2.3 on Global and Environmental Justice for further discussion).

Finally, across many discussions about algorithmic injustice, the insights that non-Western traditions can provide are worth further exploration. Birhane (2021) has used Ubuntu's relational ethics to highlight the value of relationality and to rethink how AI practices shape the web of relations. Wong (2025) has drawn from Confucian philosophy to rethink trustworthiness and trustworthy AI. Green (2025) has compared the AI ethics principles developed by tech corporations in Silicon Valley and those proposed by the Roman Catholic Church to examine different approaches and facilitate potential dialogue between technology and religion. Given AI systems' growing prevalence and global impact, the need to engage with diverse perspectives and epistemic traditions is as urgent as ever in theorizing the many faces of algorithmic (in)justice.

9.2.2 Rectificatory Justice

Rectificatory justice is broadly concerned with how to respond to injustice in its aftermath, in an effort to rectify or repair wrongful harm. While AI plausibly intersects with this topic in numerous ways, we restrict ourselves to two prominent aspects of the contemporary debate. The first focuses on issues of liability to compensation. Who is responsible for harms caused by AI? Relatedly, who incurs the duty to remedy the harm caused? If political theory is to offer adequate normative guidance to policymakers designing regulations for AI, including by rethinking tort liability for AI-inflicted harms, it must wrestle with this question. The second aspect concerns the other main way in which the state responds to wrongdoing: criminal punishment. Here the most discussed issue concerns the use of AI technologies in the administration of criminal justice.

Consider, first, the general issue of responsibility for harm, on which compensation claims turn. In contexts in which people may suffer harms caused by AI, assigning responsibility is of special moral urgency. Think, for example, of crashes involving autonomous vehicles, harms and destruction caused by autonomous weapons, or harms caused by delayed, incorrect, or insufficient treatment as a result of errors by AI tools in healthcare. Consider, too, administrative contexts in which people may be unfairly burdened or denied resources as a result of mistaken automated decision-making.

PROOFS

One prominent concern is that autonomous systems' autonomy – understood as a certain kind of independence from human agents – makes it difficult, if not impossible, to assign moral or legal responsibility to human persons for acts performed and harms inflicted by AI. Neither operators nor programmers nor manufacturers may be sufficiently directly morally or causally linked to AI-inflicted harm to be held meaningfully accountable. The influential notion of a "responsibility gap" describes this presumptive "gap" between machine action, on the one hand, and responsibility for those actions, on the other (Matthias, 2004; Sparrow, 2007). More recently, the label has come to be used to describe several different concerns about ascribing responsibility for AI-inflicted harms (Santoni de Sio & Giulio Mecacci, 2021).

One question is, "Who is 'on the hook' for harms caused by AI?" More precisely, "Under what conditions do people become liable to punishment, and under what conditions do people incur obligations to compensate victims for harms caused by AI?" Another question is whether responsibility gaps really exist and, if they do, whether they can be closed. One complication here arises from the somewhat ambiguous nature of relevant "gaps." On the one hand, the question of whether a responsibility gap is bridgeable arises only if there really exists at least a *prima facie* gap. On the other hand, some views suggest that, if a *prima facie* gap can be closed, there is no relevant gap in the first place.

Perhaps harms caused by AI can be traced back to human agency to a sufficient degree to hold persons morally responsible for those harms (Simpson & Müller, 2016). Perhaps human agents have the normative power to deliberately take responsibility, to make themselves answerable, for harms caused by AI (Kiener, 2022). Or, perhaps, we should view AI systems as physically instantiated socially constructed institutions, such that any difficulty to attribute responsibility is no different from any other collective action problem (Robillard, 2018). Perhaps the gap can thus be closed by treating institutions and group agents, rather than individual persons, as relevant responsibility-bearers (Taylor, 2021).

Not everyone assumes that responsibility gaps are necessarily objectionable. On some views, certain gaps might actually be desirable. They may be beneficial, at least in contexts of tragic moral choices, because they reduce the psychological costs of responsibility that would otherwise fall on human decision makers (Danaher, 2022b).

The notion of responsibility gaps raises the broader question of what, if any, bearing the possibility of holding someone responsible should have on the permissibility of delegating decisions to AI systems. The possibility of accountability – often used synonymously with responsibility in this context (Lechterman, 2022) – is sometimes treated as a necessary condition for the permissible use of AI. A common argument, for example, is that human agents must remain in control over autonomous systems precisely because, unlike human moral agents, AI cannot be held morally or legally accountable for its actions and decisions (Asaro, 2020).

PROOFS

If the use of AI ultimately reduces harm and leads to fewer lives lost, for example, in road traffic and armed conflict, but responsibility gaps cannot be closed, we may be left to choose between states of affairs in which less harm is caused but there is no possibility of accountability and states of affairs in which more harm is caused but responsible agents can be held to account (Eggert, 2025). One resulting question for future research in this area is how to weigh the harm-reducing benefits of AI against the importance of traditional accountability mechanisms. Further questions for future research include how accountability affects the permissibility of delegating certain potentially harmful actions and decisions to AI; how accountability should affect design choices; and, under what conditions and at what cost human agents must retain a certain degree of control over AI agents specifically for the purpose of exercising responsibility.

Now, consider the issue of criminal justice. Though traditionally the purview of legal theorists, criminal justice is now widely recognized as a fundamental concern of political philosophy. Artificial intelligence is deployed in the context of criminal justice in numerous ways. One way, not strictly concerned with rectification of crime but rather its prevention, concerns the role of AI in predictive policing (Jørgensen, 2022). Artificial intelligence is also increasingly deployed downstream, in decisions about the sentencing of offenders and about their suitability for release under parole, based on risk assessments for recidivism. The actual punishments that criminals receive could thus be determined by algorithmic tools (though some scholars would quibble about whether these predictive algorithms are sufficiently sophisticated to be called "AI").

The use of predictive technology to determine the punishments of criminals is highly controversial. Some of the controversy tracks general worries about predictive algorithms, especially if training data reflect racial biases (see Section 9.2.1 on Algorithmic Injustice). Insofar as there are duties to rectify the historical injustices of institutional racism, the use of criminal rectification measures that reinforce racist injustice seem particularly perverse, reinforcing the racist injustices that pervade the very institution of punishment (Shelby, 2022). A related concern is the phenomenon of statistical discrimination, whereby someone is treated not as an individual but rather on the basis of properties of the group of which they are a member. (For relevant background, without particular focus on AI, see Enoch et al., 2012.) Some theorists argue that predictive algorithms can be unjust, even if they do not produce any downstream harms, by virtue of involving impermissible epistemic claims (Lazar & Stone, 2024).

To the extent that punishment should be determined by backward-looking considerations of retribution or communicative condemnation, forward-looking risk assessments may seem irrelevant (Kelly, 2018). After all, crime prevention and punishment are two different domains. To the extent that one endorses an instrumental approach to the justification of punishment, based on its future effects, one will be less inclined to rule out forward-looking measures *ex ante*. Given the increasing popularity of instrumental approaches to criminal

PROOFS

punishment (e.g., Tadros, 2011), it is no surprise that there is an active research agenda exploring when risk-based sentencing algorithms can and cannot be justified. (For useful overviews, see Ryberg, 2020, 2024; Ryberg & Roberts, 2022.) Fruitfully, the question of AI has also prompted theorists to return to foundational theoretical issues, such as whether repeat offenders merit greater punishment than first-time offenders (Ewing, 2019).

9.2.3 Global and Environmental Justice

Certain practices surrounding AI systems raise issues of global justice, particularly concerning the unfair distributions of benefits and burdens between the Global North and the Global South, as well as such distributions' implications for global power dynamics. This subsection highlights specific dimensions of justice that gain heightened salience when examined through a global justice perspective, which critically accounts for the power asymmetries of the prevailing, and historically rooted, global order.

One aspect that has received extensive analysis is AI's impact on the environment. Artificial intelligence systems consume vast amounts of planetary resources and impose substantial environmental costs (Crawford & Joler, 2019). The manufacturing of AI infrastructure requires vast amounts of mineral resources. These resources are largely extracted from low- and middle-income countries and have been linked to geopolitical conflict and environmental harms in those countries (Crawford, 2021; Dauvergne, 2022). Training and using AI models further requires large amounts of energy, such as electricity and water, while energy consumption by data centers (Bender et al., 2021) exacerbates environmental degradation that disproportionately burdens already marginalized populations (Sultana, 2022; Ngcamu, 2023). While the extent of AI's energy consumption is an empirical issue, the information disclosed by tech companies remains limited. The lack of transparency impedes proper evaluation and leads to polarized attitudes toward AI's environmental impact (Luccioni et al., 2025).

Besides potential benefits of AI and short-term employment impacts affecting tech workers (see Chapter 7), another part of the story is the labor behind the development of AI systems, which marks another domain of AI's extractive process and unjust impacts. Currently, the AI supply chain relies on an *international division of digital labor* (Muldoon & Wu, 2023). The high-paid, well-regarded software engineering and AI research jobs are clustered in the Global North. But many other types of human labor behind the AI supply chain – from workers supporting the physical infrastructure of AI (e.g., miners and workers assembling electronic components) to data workers (e.g., those who undertake data preparation and annotation, verify algorithmic results, and review content for moderations) – tend to be low-paid. They also work under precarious and even harmful and exploitative conditions. And they are largely located in the rest of the world (Gray & Suri, 2019;

PROOFS

Tubaro et al., 2020; Roberts, 2019; Crawford, 2021). Even though AI supply-chain labor plays a crucial role in AI's development and deployment, corporate AI narratives and the design of microwork platforms have been found to obscure and devalue certain forms of human labor, thereby further enabling labor exploitation for the benefit of some large tech corporations (Gray & Suri, 2019; Crawford & Joler, 2019).

In addition to material injustices, concerns have been raised about AI's unjust impacts in reinforcing the Western hegemony of values and epistemologies. Currently, AI models are primarily trained on high-resource languages – those with large amounts of digitized data, such as English, French, and German (with English content being the most dominant one). While many other languages are spoken by huge populations, their countries have significantly lower quantities of digitized data. This has created a *digital language divide* in which English speakers are granted greater access to advanced AI tools than speakers of lower resource languages (Ta & Lee, 2023). Furthermore, studies reveal that Western-centric, especially American-centric, perspectives are often encoded in AI models (Johnson et al., 2022). The skewed perspectives embedded in AI systems have given rise to concerns about Western hegemony and cultural imperialism (Birhane, 2023; Muldoon & Wu, 2023). This situation also echoes what Cave and Dihal (2020) call the "Whiteness of AI" – the fact that AI is predominantly portrayed and perceived as being for and about white people, which perpetuates and exacerbates the racialized hierarchy of power.

Based on these observations, burgeoning scholarship has recently argued that the power dynamics shaped by AI systems and surrounding industrial practices mirror and further extend the extractive dynamics of colonialism (Kwet, 2019; Ricaurte, 2019; Couldry & Mejias, 2019; Coleman, 2019; Amrute, 2019; Oyedemi, 2019; Mohamed et al., 2020; Adams, 2021; Birhane, 2023; Muldoon & Wu, 2023; Nothias, 2025). While using slightly different terms – including *digital colonialism*, *data colonialism*, *algorithmic colonialization*, and *algorithmic coloniality* – the core concern is the concentration of economic and political power that big tech companies, most of which are based in the US, hold over the rest of the world. These companies conduct extractions in the name of progress, while imposing disproportionate harms and violence (including environmental degradation, labor exploitation, and cultural imperialism, as discussed earlier) on a broad range of populations, including populations of the Global South and other underserved populations, such as Black, Indigenous, and Latinx communities. Thus far, discussions have drawn on literature in decolonial studies (Maldonado-Torres, 2007; Quijano, 2007; Mignolo, 2012) to delineate the mechanisms that shape these global power dynamics (Muldoon & Wu, 2023) and to explore ways to disrupt them (Mohamed et al., 2020; Adams, 2021).

Many questions about digital colonialism remain to be further theorized. Both the concept of digital colonialism and the adequacy of related terminology

PROOFS

are issues of scholarly debate (Moosavi, 2020; Mouton & Burns, 2021; Táíwò, 2022). Further examination is also needed to historicize the idea of digital colonialism, analyze its connections to traditional forms of colonialism and coloniality, and explore potential responses to it. Last, but not least, while scholars largely converge on diagnostic aspects of digital colonialism, one area that warrants further scholarly exploration is the compatibility of their prescriptive approaches (Nothias, 2025).

9.2.4 Intergenerational Justice

According to some commentators, an already existing cluster of concerns about existential risks arising from climate change and other global threats now also includes potentially catastrophic risks posed by AI (cf. Bales et al., 2024). Rather than discuss the possibility of existential risk and the future of humanity, we will limit our discussion here to issues of intergenerational justice. Intergenerational justice concerns what duties, if any, present-day generations have to consider justice for past and future people (Meyer, 2021). Since AI creates new ways in which we in the present can exercise power over people who are no longer alive and people who do not yet exist, it is an appropriate subject of intergenerational justice.

One foundational question is whether valid justice claims exist only between people who can harm or benefit one another (Barry, 1989; Buchanan, 1990; Gauthier, 1987; Heyd, 2009; Hiskes, 2009). AI has no bearing on the general question of between whom duties exist. But to the extent that AI affects whether and how we, in the present, might harm, wrong, or benefit people who are no longer alive or who do not yet exist, AI may bear on questions of intergenerational justice. While political theorists have come to grapple with how AI mediates power relations between living people (Lazar, 2025), a nascent area of research concerns how AI mediates the power we, in the present, exercise over both people who are no longer alive and people who do not yet exist.

Consider, first, so-called afterlife technologies. Advances in generative AI have made it possible, in effect, to digitally revive the dead (Worth, 2025; Dickey, 2025; Hern, 2024; Yang, 2024). Users can now create interactive avatars and AI companions that are essentially deepfakes of deceased people. "Griefbots" or "deadbots" are mainly marketed as providing new, digital ways of grieving. But deepfakes and chatbots that mimic deceased people could, in principle, appear anywhere; including in schools or museums, allowing people to have conversations with historical figures. Anyone of whom there exist audio and video recordings that can be used to train models will be revivable as an avatar.

Setting aside reasons we might have *not* to digitally revive the dead, questions raised by afterlife technologies include the following. How might generative AI extend theories of privacy to the afterlife? What does valid consent

PROOFS

require to digitally replicate someone and, relatedly, on what should people be consulted concerning how their data may be used after their death? Finally, how should digital afterlife technologies be regulated to protect deceased people's rights and interests, and to minimize risks of psychological harm to living users of deadbots?

Now consider future generations. One characteristic of intergenerational relations is the asymmetry in power relations between those currently alive and those who are not. Whereas power relationships between contemporaries can change frequently, present generations necessarily have asymmetrical power over future generations (Meyer, 2021; Barry, 1977; Barry, 1989). Present generations can affect the circumstances of future generations. AI–related actions and policies that deplete the world's natural resources may well set back future generations' interests or reduce their options (cf. Barry, 1999; Beckerman, 1999; Heyward, 2017). Besides the circumstances in which future generations will live, present-day generations also determine which future people will exist. That future people's very identities are contingent on the actions of those who come before them has given rise to a persistent debate about the very possibility of intergenerational justice claims (Parfit, 1984). Any applications of AI that might affect what specific people will come into existence may similarly become entangled in what, following Derek Parfit, has come to be known as the non-identity problem.

One central question in debates about intergenerational justice is whether we have reasons to discount future persons' interests (Sidgwick, 1907; Rawls, 1971; Parfit, 1984; Broome, 1994; Mintz-Woo, 2021). These concerns extend well beyond questions relating to technology. Nevertheless, decisions about the development of AI models today can account for or ignore the interests of future generations. Artificial intelligence's familiar promises and perils – from significant medical advances to jeopardizing the future of humanity – also raise questions about whether AI's benefits to existing or near future people may justify risks to those in the more distant future. If AI promises to be sufficiently beneficial to present-day generations, does that justify exposing future generations to certain risks? Might powerful AI models even help determine whether, and to what extent, future people's interests may permissibly be discounted?

9.3 DEMOCRACY

Since this volume contains a dedicated chapter on Democracy/Elections (Chapter 1), our focus here will be on democratic values, rather than institutions, practices, and processes. In a remarkable example of how social (in this case, technological) developments can and should prompt us to rethink our theories (Dewey, 1927), the rise of AI has led democratic theorists to rethink values and paradigms core to the field (Bernholz et al., 2021).

PROOFS

Existing democratic governance practices involving AI have already spawned a considerable body of literature. This trend continues political theory's long-standing concern with not just governance by the state but also governance by private as well as other types of public institutions (Adams, 2018; Anderson, 2017; Kirby, 2022). While normative evaluation of governing institutions that rely on algorithms has been in the public consciousness for a while (Danaher, 2016), contemporary AI has spurred theorists to consider whether and how AI should be used to mediate relations in a democratic society (Danaher, 2022a; Lazar, 2024, 2025) and in what circumstances the use of AI entails, for example, a democratic duty of explanation (Lazar, 2024).

Another body of literature concerns the democratization of AI. To "democratize AI" can, as a descriptive matter, mean many things – from merely increasing access to AI to ensuring democratic governance of AI. Seger et al. (2023) give a helpful categorization of various descriptive uses of the phrase "democratizing AI." Considering normative questions of whether and how we ought to democratize AI, many authors helpfully examine the semantic ambiguity the phrase licenses. Among the concepts most commonly discussed (with significant overlap) are the design, development, deployment, use, and governance of AI (Himmelreich, 2023; Lin, 2024; Noorman & Swierstra, 2023). (Note, in particular, Himmelreich's argument against the democratization of AI, conceived of as ensuring more democratic governance of AI. On AI's impact on democracy, see Coeckelbergh, 2024; Jungherr, 2023; Kreps & Kriner, 2023.)

Perhaps three of the most prominent areas on which AI has had substantive effects for political theory are (1) representation, (2) formal deliberation, and (3) the digital public sphere. These three areas will be our focus. (For practical applications of AI to democratic processes, see Chapter 1.) We will begin with a note about how to categorize AI's various uses in democratic governance.

9.3.1 Conceptualizing AI Uses in Democratic Governance

Artificial intelligence is increasingly used to enhance, support, or implement democratic governance processes in both public and private institutions. These developments have sparked a growing literature in political theory concerned with when and how, if at all, AI should be incorporated into democratic governance processes (Schneier & Sanders, 2025; Boehmer et al., 2025; Devine et al., 2023; Edelman, 2023; Fish et al., 2024; Konya et al., 2023; Lazar, 2025; Mendoza, 2023; Tessler et al., 2024).

Two different conceptual schemas for categorizing different uses of AI have emerged so far. The first is the Input-Process-Output model (Han, 2025). This is a four-fold categorization that focuses on whether a given AI application in a democratic process (1) "originate[s] from the citizenry" (input) and (2) "whether AI decisions carry legal or authoritative weight" (authority of output). Han then assesses the four resulting categories of AI application

PROOFS

(democratically binding, undemocratically binding, democratically unbinding, undemocratically unbinding) in terms of democratic values: "inclusive and equal participation, quality of decisions, deliberation, and the autonomy of citizens to set the political agenda" (Han, 2025).

The second schema is Lazar and Manuali's categorization of the various ways LLMs are currently used in proposed democratic processes: *summarization*, *aggregation* of preferences, *facilitation* of deliberation, and *representation* (Lazar & Manuali, 2026). These uses are not mutually exclusive. For example, facilitation of deliberation may involve using LLMs to summarize conversation. Generalizing from Lazar and Manuali's focus on LLMs, we will examine how AI has been theorized and used in democratic governance most commonly: *representation* and *deliberation*.

9.3.2 AI in Representation

Perhaps the most prominent example of the influence of AI on democratic theory stems from efforts by computer scientists and democratic practitioners to use AI as representatives for people (Boehmer et al., 2025; Fish et al., 2024; Gudiño et al., 2024). One could apply existing theories of political representation to extract a normative analysis of such efforts.

The technological possibility that AI could predict and extrapolate our preferences over possible policy options with relative accuracy (Boehmer et al., 2025; Fish et al., 2024; Konya et al., 2022) raises the normative question of whether AI could *legitimately* represent us. Lechterman (2024) worries that the normative viability of such a proposal would reduce representation to the mere prediction of preferences. After summarizing some of the arguments in favor of digital representatives (what he terms "Avatar Democracy"), Himmelreich (2023) argues that digital representatives run into the same classic problem as human representatives. Namely, they might deliberate with other representatives and disagree with the preferences of their constituent(s) due to expertise that they have, at which point justifying and explaining their decision(s) to their constituent(s) (a requirement of legitimacy) will be difficult. Han (2025) sees promise in AI representatives to which duties by citizens can be delegated with those citizens' authorization, and which can be overseen by citizens to ensure that those AI representatives continue to act within their remit. Taylor (2025) also gives a more tentative answer in the affirmative by outlining conditions under which current or near-term-future AI could act as a representative in a normatively robust sense – though he notes that many conditions (including ones involving AI representatives continuously consulting with citizens) would have to be met for that to be the case.

But beyond the question of whether AI can legitimately represent us, the technological possibility of AI representatives also allows us to challenge some of the tenets of representation theory. Taylor's (2025) approach is especially illuminating. It adjusts classic formulations of principles in representation

PROOFS

theory in such a way that an AI representative could, assuming certain conditions and constraints, meet them. The challenge, then, is to figure out whether, and why, such revised principles are implausible, or to accept that the original formulations were mistaken. The technological possibility of AI representation thus allows us to try and refine representation theory, either by affirming or by adjusting the original formulations in light of AI-inspired challenges.

The relationship of AI to systems of representation will depend on the capabilities of AI models, of course. How accurate our predictions are with respect to the future capabilities of AI representatives (real or hypothetical) determine how we might be able to challenge current formulations in representation theory. For instance, Taylor's (2025) normative argument in favor of AI representatives depends on a suitably idealized AI representative, with a particular suite of abilities that no currently existing projects that attempt to instantiate AI representation seem to have. If theorists make assumptions that are implausible in light of present and future AI abilities, any normative proposal for an AI representative will be vulnerable to the critique that it is too remote a possibility to consider. It may also be subject to a "problem of the second best."

The rise of advanced AI has also led democratic theorists to consider, at least hypothetically, the possibility of a completely AI-run government. Lovett and Zuehl (2022) conclude that something would be missing from such a world: the noninstrumental value of collective self-determination, which they conceive of as a form of distinctly democratic autonomy. This is an example of how engaging with questions about AI may lead to a contribution in its own right to the literature relating to justifications for democratic governance. Contemporary AI has thus helped articulate a value that purely instrumentalist accounts of democracy have not been able to capture.

9.3.3 AI in Formal Democratic Deliberation

One prominent set of proposals for using AI in democratic governance involves employing AI in *democratic deliberation* (see Chapter 1). Formal deliberative processes are costly, and it is especially difficult to train and pay for human moderators. Landemore (2024) asks if AI can be used, then, to bring "quality deliberation to the masses." She considers two different proposals for scaled-up deliberation – Mass Online Deliberation and Multiple Rotating Mini-Publics – and several different roles for AI in scaling up these proposals to allow for more participation. She suggests that AI could facilitate, summarize, select participants, translate, fact-check, and aggregate data (Landemore, 2024). Still, whether AI can actually "scale up" deliberation remains to be seen. One recent study suggests that AI-based facilitation incurs a "penalty" in that people prefer human facilitators in deliberation and so participate less in AI-facilitated deliberation (Jungherr & Rauchfleisch, 2025).

Besides asking whether and how AI can "scale up" deliberation, one might ask what normative constraints should be placed on AI as we employ it in

PROOFS

formal deliberative processes, such as in mini-publics. Mikhaylovskaya suggests that AI's guiding value *qua* facilitator ought to be (democratic) equality (Mikhaylovskaya, 2024). She proposes that AI (as opposed to humans) might be better at being an impartial facilitator and at making sure all voices are included. In her view, we ought to design AI facilitators in light of the noninstrumental democratic value of equality.

Considering the use of AI in citizen assemblies, McKinney (2024) offers both (1) a conceptual framework for the various functions of AI in citizen assemblies and (2) a normative framework that allows for the assessment of these functions with respect to democratic goods.

A common thread among those focusing on incorporating AI into formal deliberative processes emerges from the original deliberative-democratic tradition dating back to Cohen and Habermas (Cohen, 1989; Habermas, 1996). Deliberative democratic practitioners in the present context often focus on the ability of AI to *identify consensus* or *agreement* (Devine et al., 2023; Konya et al., 2023; Tessler et al., 2024). However, other deliberative democrats resist the claim that deliberation should (in some sense) aim at consensus or agreement (Fraser, 1990; Young, 2002). In a similar spirit, Palomo Hernández (2025) critiques one popular project that incorporates AI into a democratic decision-making process: the Habermas Machine (Tessler et al., 2024). In addition to critiquing the lack of engagement between participants, Palomo Hernández (2025) argues that the Habermas Machine "overemphasises the desirability of agreement in deliberative processes."

Other theorists have more recently taken an even less deliberation-focused approach. Instead of proposing to use AI in formal deliberative processes (such as in mini-publics), Farrell and Han (2025) suggest that we ask how we can use AI to create, maintain, and strengthen the power of *democratic publics*. Reminiscent of the agonistic democratic tradition (Mouffe, 1993), they remind us that a crucial aspect of politics, in general, and democracy, in particular, is the negotiation of power relations between groups, and that democracies require people to recognize themselves as part of publics to tackle shared problems (Dewey, 1927). Farrell and Han (2025) thus call for both theorizing and empirical research related to the question of how AI can support the formation of publics and their negotiation of social problems.

9.3.4 The Digital Public Sphere

Consider now the place of much of the contemporary democratic action: the digital public sphere. Our public discourse is increasingly mediated by AI. We are, as Seth Lazar puts it, "connected by code" (2026). The most conspicuous way in which AI mediates our public discourse concerns its role on social media platforms. For better or worse, these platforms continue to serve as the most significant fora in which people argue with one another about important social and political issues, express their opinions, and learn information about

PROOFS

current political events. Facebook, Instagram, X, TikTok, BlueSky, YouTube, and other spaces together constitute our new digital public sphere.

How does AI govern our communications in this space? First, AI systems enforce the rules on what speech is allowed. While humans formulate the general rules prohibiting harmful content (from hate speech to death threats to child abuse material), AI systems play a fundamental role in applying and enforcing them. Second, AI systems determine what content people see. Algorithms tuned to optimize our engagement control what posts appear in our feeds. Finally, machines loom large on social media because the content we encounter is itself often synthetically produced – the product of generative AI tools powered by large language models (LLMs), such as ChatGPT, Claude, and Grok. Increasingly, chatbots are integrated into the platforms, themselves participants in our digital public conversation.

Each of these three areas raises important normative questions, including for political theorists – concerning the moral duties of both private corporations and policymakers tasked with regulating them. Consider the role of AI in content moderation: the process by which platforms govern the speech of their users. Insofar as such policies have a profound effect on users' communicative interests, they are arguably evaluable by normative standards of political theory, and when such policies are compelled by the state, they are inarguably so (see Lazar, 2026; Howard, 2026a).

One question concerns permissible and impermissible uses of AI by platforms as part of their content moderation practices. Social media platforms deploy supervised machine learning (ML) models, trained on massive data sets of labelled training data. For each type of prohibited content, platforms build what is known as an ML *classifier* to adjudicate whether a fresh piece of content is allowed or disallowed. (This approach replaces older, cruder uses of technology to moderate content – such as simple lists of banned keywords.) Suppose a user on X or Facebook offers a robust critique of a politician – perhaps one that could arguably be interpreted as a threat (see Fisher & Howard, 2024). Suppose the speech is then removed, thanks to an AI classifier. Why might the user object?

There are many potential objections – all applications of more general worries about deploying AI in decisions that impact our fundamental interests. Some have argued for the notion of a fundamental right to a human decision (for detailed treatment of this view, see Huq, 2020). Others have raised worries about the inaccuracy of AI, though there are questions as to whether humans are really "better," not least since they cannot do the relevant work at scale. In the content moderation context, AI classifiers struggle to interpret speaker intent or purpose, such as inferring sarcasm and nuance – leading to false positives (Douek, 2021). Likewise, efforts to evade rules by speakers outsmarting AI leads to false negatives. And there are familiar worries about bias, and that AI will end up producing unfairly worse results for members of historically marginalized groups (see Section 9.2.1 on Algorithmic Injustice).

PROOFS

One particularly instructive objection concerns the *opacity* of AI models – the fact that they constitute a so-called black box, such that even engineers often cannot explain fully why those models produce the verdicts that they do. Machine learning models can detect patterns in training data far more subtle and complex than humans ever could themselves. But this may make it difficult for users to understand and predict when, and why, their speech is likely to be affected – and to contest mistakes, holding power accountable. This impairs users' interest in what Kate Vredenburgh, writing about AI opacity in general, calls *informed self-advocacy* (2022). Note that, while this analysis is inspired by the case of AI, it has wider ramifications – helping us see, for example, what is objectionable about opaque bureaucratic rules that citizens cannot successfully navigate or contest. In this way, Vredenburgh's work fruitfully illuminates how attention to problems raised by AI prompts us to make progress in more fundamental normative questions. (For related analysis, see Lazar, 2024.)

These objections need not render the use of AI impermissible in the governance of online speech. Rather, they constrain what sort of AI should be used, and what requirements of transparency and oversight properly apply to it. Further, the costs of using AI must be weighed against the benefits. Content moderation is a defensive effort to prevent harm, and preventing harm requires speed. Faced with billions of user-generated posts per day, the large platforms cannot feasibly catch the surfeit of harmful content without the use of technology. Insofar as platforms have a duty to combat such content, they may have a duty to use AI (Howard, 2026a, though cf. Barnes, 2022).

What about the role of AI in *curating* public discourse, such as deciding what content users see in their recommended algorithm-powered newsfeeds? Political theorists are only beginning to address this question. As Lazar argues, settling the boundaries of permissible and impermissible online speech does not settle this problem (2026). Rather, we need a positive vision of what a flourishing public sphere ought to involve. Despite historic attention to this question by luminaries, such as John Dewey (1926), Iris Marion Young (2002), Jürgen Habermas (1989, 1996), and John Rawls (2005), there has been insufficient effort to update our conception of a public sphere for the digital age (for an early exception, albeit in legal theory, see Balkin 2007). But progress is afoot. Some work addresses the principles that should govern the amplification of content (Brown 2026; Lazar, 2026; Miller, 2021). Some work explores the ways that amplifying some (otherwise innocuous) speech can cause harm (Howard & Kira, 2026), as with forms of online public shaming (Billingham & Parr, 2020). Other work focuses on the discursive norms that citizens ought to follow in their online discourse (Cohen & Fung, 2021). And some work focuses on how to design online spaces to build community (Forestal, 2021). There is also a raft of work on such problems as polarization which, while not limited to the digital realm, may be exacerbated by it (Sunstein, 2017; Talisse,2019), as well as on the problem of manipulation (Renzo, 2025).

PROOFS

A final set of concerns about AI involves the burgeoning use of LLM-powered chatbots. Insofar as these chatbots are used as new search engines, they bring back to the fore familiar questions about how citizens ought to receive their information (Sunstein, 2017), and what the normative principles are for governing search (Grimmelmann, 2014). Indeed, the question of the appropriate role of the news media in an age of algorithmic media, and how to rebuild shared media institutions that command widespread trust, remains seriously under-researched by political theorists relative to its importance (though see Heawood & Peter, 2023) – a problem only exacerbated by LLMs. Some existing work concerns the moral obligations of AI companies that produce public-facing chatbots (or other content-creation tools, such as image or video generators) – for example, concerning how to align model outputs with important moral and political values (e.g., Kasirzadeh & Gabriel, 2023). Other work concerns the downstream obligations of platforms that host AI-generated content (Fisher et al., 2024).

Cutting across all these concerns is a research agenda on the political economy of Big Tech. Many argue that antitrust and anticompetition policy tools, or reforms to business models, are the appropriate remedy for the pathologies of our digital public sphere (Balkin, 2021, 2022; Hindman, 2018; Reich et al., 2021). In this way, concerns about AI are fruitfully reinvigorating broader debates about economic justice.

9.4 RIGHTS AND THE FUTURE

This section explores the future of rights in a world shared with AI. We will first consider the future of *human* rights in a world in which our lives increasingly depend on technology. We will then briefly consider the contested notion of *robot* rights.

9.4.1 Human Rights

Three "generations" of existing human rights are commonly distinguished: (1) civil and political rights; (2) economic, social, and cultural rights; and (3) collective rights. Artificial intelligence's transformative impact on critical areas of human life and activity has led to calls to recognize a new, fourth "generation" of human rights, fit for the digital age (von Schirach, 2021; Risse, 2023; Shany, 2023). These include a right to digital self-determination, the right that certain algorithms be "transparent, verifiable and fair," and that "major decisions" be "taken by a human being" (von Schirach, 2021) as well as rights to protect our "epistemic actorhood" as knowers and knowns and, ultimately, a right to the exercise of *human* intelligence (Risse, 2023).

No consensus exists on the need for a new generation of human rights for the digital sphere. On the one hand, a longstanding worry is that positing too many human rights devalues their currency. If we treat everything important

PROOFS

as a right, the worry goes, then rights become meaningless (Griffin, 2008). On the other hand, some commentators have argued, if human rights are to remain relevant in a changing world, the human rights framework must adapt as transformative technologies advance (Schulz & Raman, 2020). One foundational question AI brings to the fore in human rights theory, then, is how to adjust human rights to changing social norms and new technological advances while avoiding longstanding fears about conceptual pitfalls, including rights inflation, overreach, and expansionism. In short, as transformative technologies advance, should human rights evolve? If so, how?

Even if new technology requires a reassessment of human rights, the determination of which rights should make the list remains contested. As noted earlier, proponents of new human rights have invoked rights to protect people from new and distinct, AI-precipitated harms. Calls for new human rights raise the question whether nonhuman, artificial forms of agency create new and distinct ways of harming and/or wronging people. This possibility, in turn, raises the question of whether we need new legal protections for people – in particular, as subjects of algorithmic decision making. While legal rights depend on institutional recognition and enforcement, moral rights exist independently of political and legal institutions. What, if any, *moral* rights are available to humans, nonhuman animals, or nonhuman, artificial agents does not depend on what rights are *legally* recognized.

An enduring question is what "grounds," or what justifies, human rights. A prominent assumption is that all human beings have human rights simply by virtue of being human (Griffin, 2008; Simmons, 2015). But it is contested whether there is anything special about being human that can explain the existence of human rights. What feature of "being human" could possibly explain – in a manner acceptable to people with different commitments, from different cultures, and different religious backgrounds – what makes humans "special" and why we have certain rights simply by "virtue of our humanity"?

One popular response in human rights theory has been to divorce the defense of human rights from attempts to specify their foundations, and to provide an account of their function in global political practice that does not rely on some contested view about the significance of humanity in virtue of which we have human rights (Rawls, 1999; Cohen, 2004; Beitz, 2009). While human rights theorists have been divided over the significance of the notion of "our humanity," the arrival of AI on the scene might sharpen views over time about whether there is, in the end, anything special about "being human." By providing a new foil to humans, nonhuman, artificial agents may help us advance our understanding of the notion of different entities' "moral status" (Risse, 2023; Liao, 2020, 2010). They might bring into sharper relief what – if anything – makes humans relevantly unique. Or they may confront us with the sobering possibility that there is nothing special about human intelligence, or being human, after all, and that insisting otherwise is simply parochially "carbon chauvinist" (Sagan, 1973).

PROOFS

9.4.2 Robot Rights

Consider, finally, the possibility of robot rights (Coeckelbergh, 2010; Gunkel, 2018; Risse, 2023). If the features that ground rights are not unique to humans, we face the possibility that those same features ground rights of nonhuman entities, along with questions about whether those nonhuman entities can be subjects of justice. It is already a common view that we should recognize certain rights of nonhuman animals. (On the question of political membership for nonhuman animals, see Donaldson & Kymlicka, 2011.) A less common view is that we should recognize certain rights of inanimate objects, such as trees, rivers, or mountains. Instead of taking a position on the possibility of robot rights, we will note simply how much is at stake in getting the answer right. To say that AI systems are moral rights-holders would be to say that they are the kinds of entities that can be wronged. Generally speaking, the conceptual possibility of being wronged comes with the possibility of generating duties of compensation, and apology, and so on, as well as, possibly, permissions to enforce the protection of rights (Bryson et al., 2017; Basl & Bowen, 2020). Hence, to argue for AI systems' rights is, in effect, to argue for the possibility of duties to compensate AI systems, to apologize to AI systems, and perhaps even the possibility of permissions for AI systems to enforce the protection of their rights. Recognizing rights comes with serious normative commitments.

In sum, by providing a foil to humans, AI might help us make progress on the longstanding question of who has rights and why. Artificial intelligence also gives new urgency to questions about whether new technological possibilities generate the need for new human rights, and about nothing less than what, in the end, if anything at all, is "special" about being human.

9.5 CLOSING REMARKS

This chapter presented a partial snapshot of the intellectual landscape during a moment of great upheaval. We cannot meaningfully address, not even in theory, the political questions of our time without grappling with the impact technology has on people's lives (Risse, 2023). Political theory has an important role to play as we grapple with how to live with AI, and we hope to have provided a sample of starting points for future research.

Longstanding concerns about, for example, structural racism now extend to algorithmic bias and discrimination. What one thinks about AI and the future of work (though we didn't discuss this here: see Chapter 7) will likely depend at least partially on one's views about meaning in life, distributive justice, and equality of opportunity. And one's views about robot rights will likely be informed by one's views about human rights and those of nonhuman animals.

Different theoretical perspectives will continue to come to different conclusions. Transformative though it is, AI changes nothing about political theory at the foundational level. Foundational questions at the heart of political theory

PROOFS

concerning at least the nature of certain values are, in a sense, impervious to the contingencies of a changing world. What one should think about AI's appropriate place in our social and political lives may, in many cases, be determined by one's views in political theory.

BIBLIOGRAPHY

Adams, N. P. (2018). Institutional Legitimacy. *Journal of Political Philosophy*, *26*(1), 84–102.

Adams, R. (2021). Can Artificial Intelligence Be Decolonized? *Interdisciplinary Science Reviews*, *46*(1–2), 176–197. https://doi.org/10.1080/03080188.2020.1840225

Amrute, S. (2019). Tech Colonialism Today. In EPIC [Keynote Address]. EPIC2019. www.epicpeople.org/amrute-tech-colonialism-today/

Anderson, E. (1999). What Is the Point of Equality? *Ethics*, *109*(2), 287–337. https://doi.org/10.1086/233897

Anderson, E. (2017). *Private Government: How Employers Rule Our Lives (and Why We Don't Talk about It)*. Princeton University Press.

Asaro, P. (2012). On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-making. *International Review of the Red Cross*, *94*(886), 687–709. https://doi.org/10.1017/S1816383112000768

Asaro, P. (2020). Autonomous Weapons and the Ethics of Artificial Intelligence. In S. M. Liao (Ed.), *Ethics of Artificial Intelligence* (pp. 212–236). Oxford University Press.

Aytac, U. (2024). Digital Domination: Social Media and Contestatory Democracy. *Political Studies*, *72*(1), 6–25. https://doi.org/10.1177/00323217221096564

Bales, A., D'Alessandro, W., & Kirk-Giannini, Cameron Domenico. (2024). Artificial Intelligence: Arguments for Catastrophic Risk. *Philosophy Compass*, *19*(2). https://doi.org/10.1111/phc3.12964

Balkin, J. M. (2007). Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society. In P. S. Berman (Ed.), *Law and Society Approaches to Cyberspace*. Routledge. https://doi.org/10.4324/9781351154161

Balkin, J. M. (2021). How to Regulate (and Not Regulate) Social Media Symposium: Free Speech and Social Media Platform Regulation. *Journal of Free Speech Law*, *1*(1), 71–96. https://heinonline.org/HOL/P?h=hein.journals/jfspl1&i=71

Balkin, J. M. (2022). To Reform Social Media, Reform Informational Capitalism. In L. C. Bollinger & G. R. Stone (Eds.), *Social Media, Freedom of Speech, and the Future of our Democracy* (pp. 233–254). Oxford University Press. https://doi.org/10.1093/oso/9780197621080.003.0014

Barnes, M. R. (2022). Online Extremism, AI, and (Human) Content Moderation. *Feminist Philosophy Quarterly*, *8*(3/4). https://doi.org/10.5206/fpq/2022.3/4.14295

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press. https://mitpress.mit.edu/9780262048613/fairness-and-machine-learning/

Barry, B. (1977). Justice Between Generations. In J. Raz & P. M. S. Hacker (Eds.), *Law, Morality, and Society: Essays in Honour of H. L. A. Hart* (pp. 268–284). Clarendon Press. http://archive.org/details/lawmoralitysocieoooounse

PROOFS

Barry, B. (1989). *Theories of Justice: A Treatise on Social Justice, Vol. 1*. University of California Press.

Barry, B. (1999). Sustainability and Intergenerational Justice. In A. Dobson (Ed.), *Fairness and Futurity: Essays on Environmental Sustainability and Social Justice* (pp. 93–117). Oxford University Press. https://doi.org/10.1093/0198294891.003.0005

Basl, J., & Bowen, J. (2020). AI as a Moral Right-Holder. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford Handbook of Ethics of AI* (pp. 289–306). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.18

Beckerman, W. (1999). Sustainable Development and Our Obligations to Future Generations. In A. Dobson (Ed.), *Fairness and Futurity: Essays on Environmental Sustainability and Social Justice* (pp. 71–92). Oxford University Press. https://doi.org/10.1093/0198294891.003.0004

Beitz, C. R. (2009). *The Idea of Human Rights*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199572458.001.0001

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (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 (FAccT'21)* (pp. 610–623), Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

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

Benn, C., & Lazar, S. (2022). What's Wrong with Automated Influence. *Canadian Journal of Philosophy*, *52*(1), 125–148. https://doi.org/10.1017/can.2021.23

Bernholz, L., Landemore, H., & Reich, R. (Eds.). (2021). *Digital Technology and Democratic Theory*. University of Chicago Press.

Billingham, P., & Parr, T. (2020). Online Public Shaming: Virtues and Vices. *Journal of Social Philosophy*, *51*(3), 371–390. https://doi.org/10.1111/josp.12308

Binns, R. (2020). On the Apparent Conflict between Individual and Group Fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 514–524). https://doi.org/10.1145/3351095.3372864

Birhane, A. (2021). Algorithmic Injustice: A Relational Ethics Approach. *Patterns*, *2*(2), 100205. https://doi.org/10.1016/j.patter.2021.100205

Birhane, A. (2022). Power to the People? Opportunities and Challenges for Participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. https://doi.org/10.1145/3551624.3555290

Birhane, A. (2023). Algorithmic Colonization of Africa. In K. Dihal & S. Cave (Eds.), *Imagining AI: How the World Sees Intelligent Machines* (pp. 247–261). Oxford University Press.

Boehmer, N., Fish, S., & Procaccia, A. D. (2025). Generative Social Choice: The Next Generation (No. arXiv:2505.22939). *arXiv*. https://doi.org/10.48550/arXiv.2505.22939

Broome, J. (1994). Discounting the Future. *Philosophy & Public Affairs*, *23*(2), 128–156. https://doi.org/10.1111/j.1088-4963.1994.tb00008.x

Brown, É. (2026). Algorithmic Amplification and Democratic Equality. *Ethical Theory and Moral Practice*. https://philpapers.org/rec/BRODNR-7

Bryson, J. J., Diamantis, M. E., & Grant, T. D. (2017). Of, for, and by the People: The Legal Lacuna of Synthetic Persons. *Artificial Intelligence and Law*, *25*(3), 273–291. https://doi.org/10.1007/s10506-017-9214-9

PROOFS

Buchanan, A. (1990). Justice as Reciprocity versus Subject-Centered Justice. *Philosophy & Public Affairs*, *19*(3), 227–252. www.jstor.org/stable/2265395

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (pp. 77–91). https://proceedings.mlr.press/v81/buolamwini18a.html

Cave, S., & Dihal, K. (2020). The Whiteness of AI. *Philosophy & Technology*, *33*(4), 685–703. https://doi.org/10.1007/s13347-020-00415-6

Coeckelbergh, M. (2024). *Why AI Undermines Democracy and What to Do About It*. Polity.

Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology*, *12*(3), 209–221. https://doi.org/10.1007/s10676-010-9235-5

Cohen, J. (1989). Deliberation and Democratic Legitimacy. In D. Matravers & J. Pike (Eds.), *Debates in Contemporary Political Philosophy: An Anthology* (pp. 342–360). Routledge. https://philpapers.org/rec/COHDAD-2

Cohen, J. (2004). Minimalism About Human Rights: The Most We Can Hope *For? Journal of Political Philosophy*, *12*(2), 190–213. https://doi.org/10.1111/j.1467-9760.2004.00197.x

Cohen, J., & Fung, A. (2021). Democracy and the Digital Public Sphere. In L. Bernholz, H. Landemore, & R. Reich (Eds.), *Digital Technology and Democratic Theory* (pp. 23–61). University of Chicago Press. https://doi.org/10.7208/chicago/9780226748603.003.0002

Coleman, D. (2019). Digital Colonialism: The 21st Century Scramble for Africa through the Extraction and Control of User Data and the Limitations of Data Protection Laws. *Michigan Journal of Race and Law*, *24*(2), 417–439. https://doi.org/10.36643/mjrl.24.2.digital

Collins, P. H. (2002). *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. Routledge.

Costanza-Chock, S. (2018). Design Justice, A.I., and Escape from the Matrix of Domination. *Journal of Design and Science*. https://doi.org/10.21428/96c8d426

Couldry, N., & Mejias, U. A. (2019). Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject. *Television & New Media*, *20*(4), 336–349. https://doi.org/10.1177/1527476418796632

Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.

Crawford, K., & Joler, V. (2019). Anatomy of an AI System. *Virtual Creativity*, *9*(1–2), 117–120. https://doi.org/10.1386/vcr_00008_7

Creel, K., & Hellman, D. (2022). The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems. *Canadian Journal of Philosophy*, *52*(1), 26–43. https://doi.org/10.1017/can.2022.3

Crenshaw, K. (1991). Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review*, *43*(6), 1241–1299. https://doi.org/10.2307/1229039

Danaher, J. (2016). The Threat of Algocracy: Reality, Resistance and Accommodation. *Philosophy & Technology*, *29*(3), 245–268. https://doi.org/10.1007/s13347-015-0211-1

Danaher, J. (2022a). Freedom in an Age of Algocracy. In S. Vallor (Ed.), *The Oxford Handbook of Philosophy of Technology* (pp. 250–272). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190851187.013.16

PROOFS

Danaher, J. (2022b). Tragic Choices and the Virtue of Techno-Responsibility Gaps. *Philosophy & Technology*, *35*(2), 26. https://doi.org/10.1007/s13347-022-00519-1

Dauvergne, P. (2022). Is Artificial Intelligence Greening Global Supply Chains? Exposing the Political Economy of Environmental Costs. *Review of International Political Economy*, *29*(3), 696–718. https://doi.org/10.1080/09692290.2020.1814381

Devine, F., Krasodomski-Jones, A., Miller, C., et al. (2023). Recursive Public: Piloting Connected Democratic Engagement with AI Governance. *Recursive Public*. https://vtaiwan-openai-2023.vercel.app/Report_%20Recursive%20Public.pdf

Dewey, J. (1927). The Public and Its Problems: An Essay in Political Inquiry. *Alan Swallow*. www.jstor.org/stable/10.5325/j.ctt7v1gh

Dickey, C. (2025). A.I. Griefbots Are Just Our Latest Attempt to Talk to the Dead. *The New York Times*. www.nytimes.com/2025/07/10/style/oodeath-spiritualism-talking-to-dead.html

D'Ignazio, C., & Klein, L. F. (2020). *Data Feminism*. MIT Press. https://doi.org/10.7551/mitpress/11805.001.0001

Donaldson, S., & Kymlicka, W. (2011). *Zoopolis: A Political Theory of Animal Rights*. Oxford University Press.

Douek, E. (2021). Governing Online Speech: From "Posts-as-Trumps" to Proportionality and Probability. *Columbia Law Review*, *121*(3), 759–834. www.jstor.org/stable/27007631

Edelman, J. (2023, August 29). Democratic Fine-Tuning. *Lesswrong*. www.lesswrong.com/posts/ncb2ycEB3ymNqzs93/democratic-fine-tuning

Eggert, L. (2025). Autonomised Harming. *Philosophical Studies*, *182*(1), 1–24. https://doi.org/10.1007/s11098-023-01990-y

Enoch, D., Spectre, L., & Fisher, T. (2012). Statistical Evidence, Sensitivity, and the Legal Value of Knowledge. *Philosophy & Public Affairs*, *40*(3), 197–224. https://doi.org/10.1111/papa.12000

Estlund, D. (2012). Introduction. In D. Estlund (Ed.), *The Oxford Handbook of Political Philosophy* (pp. 3–20). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195376692.013.0000

Ewing, B. (2019). Prior Convictions as Moral Opportunities. *American Journal of Criminal Law*, *46*(2), 283–332. https://heinonline.org/HOL/P?h=hein.journals/ajcl46&i=291

Falbo, A., & LaCroix, T. (2022). Est-ce que Vous Compute? Code-Switching, Cultural Identity, and AI. *Feminist Philosophy Quarterly*, *8*(3/4). https://doi.org/10.5206/fpq/2022.3/4.14264

Farrell, H., & Han, H. (2025, August 1). AI and Democratic Publics. *Knight First Amendment Institute*. http://knightcolumbia.org/content/ai-and-democratic-publics

Fish, S., Gölz, P., Parkes, D. C., et al. (2024). Generative Social Choice. In *Proceedings of the 25th ACM Conference on Economics and Computation* (p. 985). https://doi.org/10.1145/3670865.3673547

Fisher, S. A., & Howard, J. W. (2024). Ambiguous Threats: "Death to" Statements and the Moderation of Online Speech Acts. *Journal of Ethics and Social Philosophy*, *28*(2), 208–229. https://heinonline.org/HOL/P?h=hein.journals/jetshy28&i=216

PROOFS

Fisher, S. A., Howard, J. W., & Kira, B. (2024). Moderating Synthetic Content: The Challenge of Generative AI. *Philosophy & Technology*, *37*(4), 133. https://doi.org/10.1007/s13347-024-00818-9

Forestal, J. (2021). *Designing for Democracy: How to Build Community in Digital Environments*. Oxford University Press.

Fraser, N. (1990). Rethinking the Public Sphere: A Contribution to the Critique of Actually Existing Democracy. *Social Text*, *25/26*, 56–80. https://doi.org/10.2307/466240

Fraser, N. (1995). From Redistribution to Recognition? Dilemmas of Justice in a "Post-Socialist" Age. *New Left Review*, *I/212*, 68–93.

Gabriel, I. (2022). Toward a Theory of Justice for Artificial Intelligence. *Daedalus*, *151*(2), 218–231. https://doi.org/10.1162/daed_a_01911

Gauthier, D. (1987). *Morals by Agreement*. Oxford University Press. https://doi.org/10.1093/0198249926.001.0001

Green, B. (2025). Comparative AI Ethics between Silicon Valley and the Vatican: Divergent Foundations, Convergent Initiatives, and "How-to" Ideas for Discussing and Developing Technology Ethics. In Levi Checketts & Benedict S. B. Chan (Eds.), *Social and Ethical Considerations of AI in East Asia and Beyond* (pp. 51–74). Springer Nature Switzerland.

Gray, M. L., & Suri, S. (2019). *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Harper Business.

Griffin, J. (2008). *On Human Rights*. Oxford University Press.

Grimmelmann, J. (2014). Speech Engines. *Minnesota Law Review*, *98*(3), 868. https://doi.org/10.24926/265535.1199

Gudiño, J. F., Grandi, U., & Hidalgo, C. (2024). Large Language Models (LLMs) as Agents for Augmented Democracy. *Philosophical Transactions of the Royal Society Series A: Mathematical, Physical and Engineering Sciences*, *382*(2285), 20240100. https://doi.org/10.1098/rsta.2024.0100

Gunkel, D. J. (2018). *Robot Rights*. The MIT Press.

Habermas, J. (1989). *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*. The MIT Press.

Habermas, J. (1996). *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. Polity.

Han, S. J. (2025). The Question of AI and Democracy: Four Categories of AI Governance. *Philosophy & Technology*, *38*(2), 1–26. https://doi.org/10.1007/s13347-025-00904-6

Heawood, J., & Peter, F. (2023). Political Legitimacy and the News Media: Four Normative Models of the Political Role of the News Media. In C. Fox & J. Saunders (Eds.), *The Routledge Handbook of Philosophy and Media Ethics* (pp. 83–95). Routledge. https://doi.org/10.4324/9781003134749-9

Hedden, B. (2021). On Statistical Criteria of Algorithmic Fairness. *Philosophy & Public Affairs*, *49*(2), 209–231. https://doi.org/10.1111/papa.12189

Hellman, D. (2025). Algorithmic Fairness. In Edward N. Zalta & Uri Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2025 ed.). https://plato.stanford.edu/archives/fall2025/entries/algorithmic-fairness/

Hern, A. (2024). Digital Recreations of Dead People Need Urgent Regulation, AI Ethicists Say. *The Guardian*. https://tinyurl.com/43peuze7

PROOFS

Heyd, D. (2009). A Value or an Obligation? Rawls on Justice to Future Generations. In A. Gosseries & L. H. Meyer (Eds.), *Intergenerational Justice* (pp. 167–188). Oxford University Press.

Heyward, C. (2017). Ethics and Climate Adaptation. In S. M. Gardiner & A. Thompson (Eds.), *The Oxford Handbook of Environmental Ethics* (pp. 474–486). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199941339.013.42

Himmelreich, J. (2023). Against "Democratizing AI." *AI & Society*, *38*(4), 1333–1346. https://doi.org/10.1007/s00146-021-01357-z

Himmelreich, J. (2023). Should We Automate Democracy? In Carissa Véliz (Ed.), *The Oxford Handbook of Digital Ethics*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198857815.013.33

Hindman, M. (2018). *The Internet Trap: How the Digital Economy Builds Monopolies and Undermines Democracy*. Princeton University Press.

Hiskes, R. P. (2009). *The Human Right to a Green Future: Environmental Rights and Intergenerational Justice*. Cambridge University Press.

Hoffmann, A. L. (2019). Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse. *Information, Communication & Society*, *22*(7), 900–915. https://doi.org/10.1080/1369118X.2019.1573912

Howard, J. W. (2024). The Ethics of Social Media: Why Content Moderation is a Moral Duty. *Journal of Practical Ethics*, *11*(2), 33–52. https://doi.org/10.3998/jpe.6195

Howard, J. W. (2026a). Moderation by Machine. In E. Heinze, N. Alkiviadou, T. Herrenberg, S. Parmar, & I. Tourkochoriti (Eds.), *Oxford Handbook on Hate Speech*. Oxford University Press.

Howard, J. W. (2026b). *Setting Fire to Reason: The Ethics of Free Speech*. Princeton University Press.

Huq, A. Z. (2020). A Right to a Human Decision. *Virginia Law Review*, *106*(3), 611–688. https://virginialawreview.org/articles/right-human-decision/

Johnson, R. L., Pistilli, G., Menédez-González, N., et al. (2022). The Ghost in the Machine Has an American Accent: Value Conflict in GPT-3 (No. arXiv:2203.07785). *arXiv*. https://doi.org/10.48550/arXiv.2203.07785

Jørgensen, R. (2022). Algorithms and the Individual in Criminal Law. *Canadian Journal of Philosophy*, *52*(1), 61–77. https://doi.org/10.1017/can.2021.28

Jørgensen, R. (2025). Protect, Serve, Predict? *Politics, Philosophy and Economics*. https://doi.org/10.1177/1470594X251379074

Jungherr, A., & Rauchfleisch, A. (2025). Artificial Intelligence in Deliberation: The AI Penalty and the Emergence of a New Deliberative Divide. *Government Information Quarterly*, *42*(4), 102079. https://doi.org/10.1016/j.giq.2025.102079

Jungherr, A. (2023). Artificial Intelligence and Democracy: A Conceptual Framework. *Social Media + Society*, *9*(3). https://doi.org/10.1177/20563051231186353

Kalluri, P. (2020). Don't Ask If Artificial Intelligence Is Good or Fair, Ask How It Shifts Power. *Nature*, *583*(7815), 169–169. https://doi.org/10.1038/d41586-020-02003-2

Kasirzadeh, A. (2022). Algorithmic Fairness and Structural Injustice: Insights from Feminist Political Philosophy. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 349–356). https://doi.org/10.1145/3514094.3534188

Kasirzadeh, A., & Gabriel, I. (2023). In Conversation with Artificial Intelligence: Aligning language Models with Human Values. *Philosophy & Technology*, *36*(2), 27. https://doi.org/10.1007/s13347-023-00606-x

PROOFS

Kay, J., Kasirzadeh, A., & Mohamed, S. (2025). Epistemic Injustice in Generative AI. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 684–697). AAAI Press. https://dl.acm.org/doi/10.5555/3716662.3716722

Kelly, E. I. (2018). *The Limits of Blame: Rethinking Punishment and Responsibility*. Harvard University Press.

Kiener, M. (2022). Can We Bridge AI's Responsibility Gap at Will? *Ethical Theory and Moral Practice*, *25*(4), 575–593. https://doi.org/10.1007/s10677-022-10313-9

Kira, B., & Howard, J. W. (2026). Reduce or Remove: Demotion, Content Moderation, and Human Rights. *Law & Philosophy*. https://link.springer.com/article/10.1007/s10982-026-09557-8.

Kirby, N. (2022). Institutional Integrity: Its Meaning and Value. *Ethical Theory and Moral Practice*, *25*(5), 809–834. https://doi.org/10.1007/s10677-022-10330-8

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. https://doi.org/10.48550/ARXIV.1609.05807

Kong, Y. (2026). What Is the Point of Equality in Machine Learning Fairness? Beyond Equality of Opportunity. *ACM Journal on Responsible Computing*, *3*(1), 1–25. https://doi.org/10.1145/3766539

Konya, A., Qiu, Y. L., Varga, M. P., & Ovadya, A. (2022). Elicitation Inference Optimization for Multi-Principal-Agent Alignment. https://openreview.net/forum?id=dz8i-yzXeVg

Konya, A., Schirch, L., Irwin, C., & Ovadya, A. (2023). Democratic Policy Development using Collective Dialogues and AI (No. arXiv:2311.02242). *arXiv*. https://doi.org/10.48550/arXiv.2311.02242

Kreps, S., & Kriner, D. (2023). How AI Threatens Democracy. *Journal of Democracy*, *34*(4), 122–131.

Kwet, M. (2019). Digital Colonialism: US Empire and the New Imperialism in the Global South. *Race & Class*, *60*(4), 3–26. https://doi.org/10.1177/0306396818823172

Landemore, H. (2024). Can Artificial Intelligence Bring Deliberation to the Masses? In R. Chang & A. Srinivasan (Eds.), *Conversations in Philosophy, Law, and Politics* (pp. 36–69). Oxford University Press. https://doi.org/10.1093/oso/9780198864523.003.0003

Lazar, S. (2024). Legitimacy, Authority, and Democratic Duties of Explanation. In D. Sobel & S. Wall (Eds.), *Oxford Studies in Political Philosophy* (Vol. 10, pp. 28–56). Oxford University Press. https://doi.org/10.1093/oso/9780198909460.003.0002

Lazar, S. (2025). Governing the Algorithmic City. *Philosophy & Public Affairs*, *53*(2), 102–168. https://doi.org/10.1111/papa.12279

Lazar, S. (2026). *Connected by Code: How AI Structures and Governs the Ways We Relate*. Oxford University Press.

Lazar, S., & Manuali, L. (2026). Using LLMs to Enhance Democracy. *Minds & Machines*, *36*, 12. https://doi.org/10.1007/s11023-026-09767-y

Lazar, S., & Stone, J. (2024). On the Site of Predictive Justice. *Noûs*, *58*(3), 730–754. https://doi.org/10.1111/nous.12477

Le Bui, M., & Noble, S. U. (2020). We're Missing a Moral Framework of Justice in Artificial Intelligence: On the Limits, Failings, and Ethics of Fairness. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford Handbook of Ethics of AI*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.9

Lechterman, T. M. (2022). The Concept of Accountability in AI Ethics and Governance. In J. B. Bullock, Y.-C. Chen, J. Himmelreich, V. M. Hudson,

PROOFS

A. Korinek, M. M. Young, & B. Zhang (Eds.), *The Oxford Handbook of AI Governance* (pp. 164–182). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197579329.013.10

Lechterman, T. M. (2024). The Perfect Politician. In D. Edmonds (Ed.), *AI Morality* (pp. 53–63). Oxford University Press. https://doi.org/10.1093/oso/9780198876434.003.0006

Leopold, D., & Stears, M. (Eds.) (2008). *Political Theory: Methods and Approaches*. Oxford University Press. https://doi.org/10.1093/oso/9780199230082.001.0001

Liao, S. M. (2020). The Moral Status and Rights of Artificial Intelligence. In S. M. Liao (Ed.), *Ethics of Artificial Intelligence* (pp. 480–504). Oxford University Press. https://doi.org/10.1093/oso/9780190905033.003.0018

Liao, S. M. (2010). The Basis of Human Moral Status. *Journal of Moral Philosophy*, *7*(2), 159–179. https://doi.org/10.1163/174552409X12567397529106

Lin, T.-A. (2024). "Democratizing AI" and the Concern of Algorithmic Injustice. *Philosophy & Technology*, *37*(3), 103. https://doi.org/10.1007/s13347-024-00792-2

Lin, T.-A., & Chen, P.-H. C. (2022). Artificial Intelligence in a Structurally Unjust Society. *Feminist Philosophy Quarterly*, *8*(3/4). https://doi.org/10.5206/fpq/2022.3/4.14191

Lovett, A., & Zuehl, J. (2022). The Possibility of Democratic Autonomy. *Philosophy & Public Affairs*, *50*(4), 467–498. https://doi.org/10.1111/papa.12223

Luccioni, A. S., Strubell, E., & Crawford, K. (2025, June). From Efficiency Gains to Rebound Effects: The Problem of Jevons' Paradox in AI's Polarized Environmental Debate. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (pp. 76–88). https://doi.org/10.1145/3715275.3732007

Maldonado-Torres, N. (2007). On the Coloniality of Being: Contributions to the Development of a Concept. *Cultural Studies*, *21*(2–3), 240–270. https://doi.org/10.1080/09502380601162548

Matthias, A. (2004). The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata. *Ethics and Information Technology*, *6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

McClure, E., & Wald, B. (2022). Algorithmic Microaggressions. *Feminist Philosophy Quarterly*, *8*(3/4). https://doi.org/10.5206/fpq/2022.3/4.14276

McKinney, S. (2024). Integrating Artificial Intelligence into Citizens' Assemblies: Benefits, Concerns and Future Pathways. *Journal of Deliberative Democracy*, *20*(1). https://doi.org/10.16997/jdd.1556

Mendoza, G. B. (2023, October 31). Can We Use AI to Enrich Democratic Consultations? *Rappler*. www.rappler.com/technology/features/generative-ai-use-enriching-democratic-consultations/

Meyer, L. (2021). Intergenerational Justice. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021). https://plato.stanford.edu/archives/sum2021/entries/justice-intergenerational/

Mignolo, W. D. (2012). *Local Histories/Global Designs: Coloniality, Subaltern Knowledges, and Border Thinking*. Princeton University Press.

Mikhaylovskaya, A. (2024). Enhancing Deliberation with Digital Democratic Innovations. *Philosophy & Technology*, *37*(1), 3. https://doi.org/10.1007/s13347-023-00692-x

PROOFS

Milano, S., & Prunkl, C. (2025). Algorithmic Profiling as a Source of Hermeneutical Injustice. *Philosophical Studies*, *182*(1), 185–203. https://doi.org/10.1007/s11098-023-02095-2

Miller, D. (2022). Doing Political Philosophy. In D. Butt, S. Fine, & Z. Stemplowska. (Eds.), *Political Philosophy, Here and Now: Essays in Honour of David Miller* (pp. 232–248). https://doi.org/10.1093/oso/9780198807834.003.0015

Miller, E. (2021). Amplified Speech. *Cardozo Law Review*, *43*(1). https://papers.ssrn.com/abstract=4014849

Mintz-Woo, K. (2021). A Philosopher's Guide to Discounting. In M. Budolfson, T. McPherson, & D. Plunkett (Eds.), *Philosophy and Climate Change* (pp. 90–110). Oxford University Press. https://doi.org/10.1093/oso/9780198796282.003.0005

Mohamed, S., Png, M.-T., & Isaac, W. (2020). Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology*, *33*(4), 659–684. https://doi.org/10.1007/s13347-020-00405-8

Mollema, W. J. T. (2025). A Taxonomy of Epistemic Injustice in the Context of AI and the Case for Generative Hermeneutical Erasure. *AI and Ethics*, *5*(5), 5535–5555. https://doi.org/10.1007/s43681-025-00801-w

Moosavi, L. (2020). The Decolonial Bandwagon and the Dangers of Intellectual Decolonisation. *International Review of Sociology*, *30*(2), 332–354. https://doi.org/10.1080/03906701.2020.1776919

Mouffe, C. (1993). *The Return of the Political*. Verso Books.

Mouton, M., & Burns, R. (2021). (Digital) Neo-colonialism in the Smart City. *Regional Studies*, *55*(12), 1890–1901. https://doi.org/10.1080/00343404.2021.1915974

Muldoon, J., & Wu, B. A. (2023). Artificial Intelligence in the Colonial Matrix of Power. *Philosophy & Technology*, *36*(4), 80. https://doi.org/10.1007/s13347-023-00687-8

Narayanan, A. (2018). Translation Tutorial: 21 Fairness Definitions and Their Politics [Tutorial]. Retrieved October 3, 2023, from https://facctconference.org/static/tutorials/narayanan-21defs18.pdf

Ngcamu, B. S. (2023). Climate Change Effects on Vulnerable Populations in the Global South: A Systematic Review. *Natural Hazards*, *118*(2), 977–991. https://doi.org/10.1007/s11069-023-06070-2

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

Noorman, M., & Swierstra, T. (2023). Democratizing AI from a Sociotechnical Perspective. *Minds and Machines*, *33*(4), 563–586. https://doi.org/10.1007/s11023-023-09651-z

Nothias, T. (2025). An Intellectual History of Digital Colonialism. *Journal of Communication*, *75*(5), 385–397. https://doi.org/10.1093/joc/jqaf003

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

O'Neil, C. (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

Oyedemi, T. D. (2019). Global Digital Capitalism: Mark Zuckerberg in Lagos and the Political Economy of Facebook in Africa. *International Journal of Communication*, *13*, 17–17. https://ijoc.org/index.php/ijoc/article/view/8774

PROOFS

Oyedemi, T. D. (2021). Digital Coloniality and "Next Billion Users": The Political Economy of Google Station in Nigeria. *Information, Communication & Society*, *24*(3), 329–343. https://doi.org/10.1080/1369118X.2020.1804982

Palomo Hernández, N. (2025). Towards Automating Deliberation? The Idea of Deliberative Democracy Embedded in Google's Habermas Machine. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, *8*(2), 1951–1960. https://doi.org/10.1609/aies.v8i2.36687

Parfit, D. (1984). *Reasons and Persons*. Oxford University Press. https://academic.oup.com/book/12484/chapter-abstract/163161038?redirectedFrom=fulltext

Quijano, A. (2007). Coloniality and Modernity/Rationality. *Cultural Studies*, *21*(2–3), 168–178. https://doi.org/10.1080/09502380601164353

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

Rawls, J. (1999). *The Law of Peoples* Harvard University Press. https://doi.org/10.2307/j.ctv1pncngc

Rawls, J. (2005). *Political Liberalism* (Expanded edition). Columbia University Press.

Reich, R., Sahami, M., & Weinstein, J. M. (2021). *System Error: Where Big Tech Went Wrong and How We Can Reboot*. Harper.

Renzo, M. (2025). Manipulation and Practical Agency. *Free & Equal: A Journal of Ethics and Public Affairs*, *1*(1). https://doi.org/10.16995/fe.17722

Ricaurte, P. (2019). Data Epistemologies, The Coloniality of Power, and Resistance. *Television & New Media*, *20*(4). https://doi.org/10.1177/1527476419831640

Risse, M. (2023). *Political Theory of the Digital Age: Where Artificial Intelligence Might Take Us*. Cambridge University Press. https://doi.org/10.1017/9781009255189

Roberts, S. T. (2019). *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press.

Robillard, M. (2018). No Such Thing as Killer Robots. *Journal of Applied Philosophy*, *35*(4), 705–717. https://doi.org/10.1111/japp.12274

Ryberg, J. (2020). Risk-Based Sentencing and Predictive Accuracy. *Ethical Theory and Moral Practice*, *23*(1), 251–263. www.jstor.org/stable/45294785

Ryberg, J. (2024). Criminal Justice and Artificial Intelligence: How Should We Assess the Performance of Sentencing Algorithms. *Philosophy & Technology*, *37*(1). https://doi.org/10.1007/s13347-024-00694-3

Ryberg, J., & Roberts, J. V. (2022). *Sentencing and Artificial Intelligence*. Oxford University Press.

Sagan, C. (1973). *The Cosmic Connection*. Anchor Press/Doubleday.

Sahebi, S., & Formosa, P. (2024). Artificial Intelligence (AI) and Global Justice. *Minds and Machines*, *35*(1), 4. https://doi.org/10.1007/s11023-024-09708-7

Santoni de Sio, F., & Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy & Technology*, *34*(4), 1057–1084. https://doi.org/10.1007/s13347-021-00450-x

Scheier, B. and Sanders, N. E. (2025). *Rewiring Democracy: How AI Will Transform Our Politics, Government, and Citizenship*. MIT Press.

Schulz, W. F., & Ramen, S. (2020). *The Coming Good Society: Why New Realities Demand New Rights*. Harvard University Press.

Seger, E., Ovadya, A., Siddarth, D., Garfinkel, B., & Dafoe, A. (2023). Democratising AI: Multiple Meanings, Goals, and Methods. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 715–722). https://doi.org/10.1145/3600211.3604693

PROOFS

Shany, Y. (2023). Digital Rights and the Outer Limits of International Human Rights Law. *German Law Journal*, *24*(3), 461–472. https://doi.org/10.1017/glj.2023.35
Shelby, T. (2022). *The Idea of Prison Abolition*. Princeton University Press. https://doi.org/10.2307/j.ctv2ks6t26
Sidgwick, H. (1907). *The Methods of Ethics*. Macmillan.
Simmons, A. J. (2015). Human Rights, Natural Rights, and Human Dignity. In R. Cruft, S. M. Liao, & M. Renzo (Eds.), *Philosophical Foundations of Human Rights* (pp. 138–152). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199688623.003.0007
Simpson, T. W., & Müller, V. C. (2016). Just War and Robots' Killings. *The Philosophical Quarterly*, *66*(263), 302–322. https://doi.org/10.1093/pq/pqv075
Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy*, *24*(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x
Stemplowska, Z., & Swift, A. (2012). Ideal and Nonideal Theory. In D. Estlund (Ed.), *The Oxford Handbook of Political Philosophy* (pp. 373–390). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195376692.013.0020
Stewart, H., Cichocki, E., & McLeod, C. (2022). A Perfect Storm for Epistemic Injustice: Algorithmic Targeting and Sorting on Social Media. *Feminist Philosophy Quarterly*, *8*(3/4). https://doi.org/10.5206/fpq/2022.3/4.14291
Sultana, F. (2022). The Unbearable Heaviness of Climate Coloniality. *Political Geography*, *99*, 102638. https://doi.org/10.1016/j.polgeo.2022.102638
Sunstein, C. R. (2017). *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press.
Ta, R., & Lee, N. (2023). How Language Gaps Constrain Generative AI Development. *International Journal of Comparative Studies in International Relations and Development*, *9*, 48–52. https://doi.org/10.48028/iiprds/ijcsird/ijcsird.v9.i1.03
Tadros, V. (2011). Punishment and Duty. In V. Tadros (Ed.), *The Ends of Harm: The Moral Foundations of Criminal Law* (pp. 264–292). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199554423.003.0012
Táíwò, O. (2022). *Against Decolonisation: Taking African Agency Seriously*. Hurst.
Talisse, R. B. (2019). *Overdoing Democracy: Why We Must Put Politics in its Place*. Oxford University Press. https://doi.org/10.1093/oso/9780190924195.001.0001
Taylor, I. (2021). Who Is Responsible for Killer Robots? Autonomous Weapons, Group Agency, and the Military-Industrial Complex. *Journal of Applied Philosophy*, *38*(2), 320–334. https://doi.org/10.1111/japp.12469
Taylor, I. (2025). Representative Robots: Can AI Systems Act in Our Name? *Free & Equal: A Journal of Ethics and Public Affairs*, *1*(1), Article 1. https://doi.org/10.16995/fe.18626
Tessler, M. H., Bakker, M. A., Jarrett, D. et al. (2024). AI Can Help Humans find Common Ground in Democratic Deliberation. *Science*, *386*(6719), eadq2852. https://doi.org/10.1126/science.adq2852
Tigard, D. W. (2021). There Is No Techno-Responsibility Gap. *Philosophy & Technology*, *34*(3), 589–607. https://doi.org/10.1007/s13347-020-00414-7
Tubaro, P., Casilli, A. A., & Coville, M. (2020). The Trainer, the Verifier, the Imitator: Three Ways in Which Human Platform Workers Support Artificial Intelligence. *Big Data & Society*, *7*(1), 2053951720919776. https://doi.org/10.1177/2053951720919776

PROOFS

Valentini, L. (2012). Ideal vs. Non-ideal Theory: A Conceptual Map. *Philosophy Compass*, *7*(9), 654–664. https://doi.org/10.1111/j.1747-9991.2012.00500.x

von Schirach, F. (2021). *Jeder Mensch*. Luchterhand Literaturverlag.

Vredenburgh, K. (2022). The Right to Explanation. *Journal of Political Philosophy*, *30*(2), 209–229.

Wong, P. (2025). Confucian "Trustworthy AI": Diversifying a Keyword in the Ethics of AI and Governance. In Levi Checketts & Benedict S. B. Chan (Eds.), *Social and Ethical Considerations of AI in East Asia and Beyond* (pp. 3–14). Springer Nature Switzerland.

Worth, T. (2025). Ready or Not, the Digital Afterlife Is Here. *Nature*. www.nature.com/articles/d41586-025-02940-w

Yang, Z. (2024). Deepfakes of Your Dead Loved Ones Are a Booming Chinese Business. *MIT Technology Review*. www.technologyreview.com/2024/05/07/1092116/deepfakes-dead-chinese-business-grief/

Young, I. M. (2002). *Inclusion and Democracy* (1st ed.). Oxford University Press.

Young, I. M. (2010). *Responsibility for Justice*. Oxford University Press.

Zimmermann, A., & Lee-Stronach, C. (2022). Proceed with Caution. *Canadian Journal of Philosophy*, *52*(1), 6–25. https://doi.org/10.1017/can.2021.17

PROOFS

10

AI and Research Methods

Christopher Barrie, Lisa P. Argyle, James Bisbee, Michael Heseltine, Christopher Lucas, Jon Mellon, Alexis Palmer, Margaret Roberts, and Arthur Spirling

PROOFS

10.1 INTRODUCTION

In this chapter, we survey current uses of artificial intelligence (AI) in political science and consider how these tools are reshaping political methodology. A brief note on terminology is useful at the outset. Throughout, we use "AI" primarily to refer to large language models (LLMs), including models developed by providers such as OpenAI, Anthropic, Meta, and DeepSeek.

To orient the reader, the chapter's main contributions fall into six substantive domains, each of which highlights a distinct way that LLMs are beginning to reshape political science research. Section 10.2 examines annotation and measurement, showing how LLMs can be used to classify and extract information from unstructured political data, while also arguing that their use requires new approaches to validation when human coding is no longer an unquestioned gold standard.

Section 10.3 turns to treatment generation, not only demonstrating how LLMs can expand the design space of experiments by making it easier to generate and tailor stimuli but also emphasizing the need to guard against bundled treatments and to validate construct fidelity.

Section 10.4 addresses silicon sampling, or the use of LLMs to generate synthetic proxies for human responses, and shows that the promise of this approach depends fundamentally on matching validation strategies to the inferential purpose of the synthetic data.

Section 10.5 considers generative agent-based modeling, explaining how LLMs can replace hand-coded behavioral rules with richer, language-mediated

Some authors of this chapter used AI tools, including OpenA GPT 5.3, Codex 5.2, and Claude Code Opus 4.6 to: identify typographic and flow errors, search for citations, reformat structure, redraft sections, and draft initial sections. Each author or author pair was responsible for their respective sections. Christopher Barrie wrote the Introduction and Discussion as well as sections 10.5 and 10.8; Jon Mellon and Margaret Roberts wrote section 10.2; Christopher Lucas and Alexis Palmer wrote section 10.3; Lisa Argyle wrote section 10.4; Michael Heseltine wrote section 10.6; and James Bisbee and Arthur Spirling wrote section 10.7.

Tucker_9781047774734_C010.indd 313 07-Oct-26 6:35:54 AM

agent policies, while also introducing new challenges of specification, interpretability, and empirical validation.

Section 10.6 examines tool-augmented data collection, focusing on the use of LLMs as research agents that can search, retrieve, parse, and organize information across heterogeneous sources. Its central contribution is to show how these systems can lower the cost of data assembly while making provenance, logging, and source verification newly important.

Section 10.7 then shifts perspective and treats LLMs not as instruments but as objects of substantive inquiry. It outlines how political scientists can audit and evaluate models themselves, whether to understand their usefulness as research tools, their social and political effects, or their broader implications for how we think about agency, reasoning, and bias.

Each section is written by one or two contributors selected for their expertise in the relevant domain. Unsurprisingly, the authors do not agree on every aspect of how AI should be used in political science. That lack of full consensus reflects the present moment: one of rapid change, methodological experimentation, and genuine uncertainty about the consequences of these emerging technologies. This lack of consensus is perhaps most obvious when it comes to questions of replication standards when using LLMs and the implications of LLMs for transparency and reproducibility more generally. Drawing on Sections 10.2–10.7, we identify both the emerging points of agreement and the open methodological questions that remain.

10.2 LLMs for Annotation and Measurement

10.2.1 The Task

Political science studies people, bureaucracies, and governments that produce large quantities of texts, images, and videos. From political speeches to bureaucratic agreements to social media videos, political scientists often need to distill and process information from unstructured data. While historically political scientists would distill these texts, images, and videos themselves, with the help of research assistants or crowd-sourcing, increasingly this processing can be done more efficiently and cheaply by AI.

In this chapter, we explore both the promise and pitfalls of using LLMs to label or measure something from unstructured data for further downstream analysis. The task we take on in this chapter is simple: A political scientist wants to take a dataset of images, texts, or videos and assign them to categorical or numerical values. They then want to use the outputs in downstream analysis, for example, to describe the whole dataset or to connect the extracted measure with other metadata.

We emphasize that using LLMs for this task still requires human validation and, in what follows, we give guidance on the choices researchers must make while using LLMs for annotating and for deciding how best to validate their

PROOFS

accuracy. In addition, we surface considerations – from personalization to replication to training data – that are unique to LLMs and merit further discussion in the field.

10.2.2 Where We Came From

Labeling with LLMs has two distinct lineages: human research assistants and traditional machine learning and statistical methods. Since the start of political science, researchers have wanted to convert unstructured text such as manifestos or political speeches into quantifiable categories. For instance, Lasswell (1941) coded newspaper headlines by whether they mentioned the US. For most tasks this has simply involved having human beings (either the researcher themselves or their students) classify the text into categories. Here, researchers are expected to have a clear definition of their measure, good coding rules, and high intercoder reliability among human labelers (Neuendorf 2016).

While substantial work has gone into trying to make such labels reliable, the baseline reproducibility across humans is not typically high, with Kappa scores of 0.8 being referred to as "almost perfect" despite meaning that 10 percent of cases would be coded differently (for a binary metric with prevalence of 50 percent). Against this backdrop, LLM use cases that replace human workflows often seem like an improvement in reproducibility.

More recently, human labelers have been used as training datasets for machine-learning algorithms and classifiers to more systematically label larger sets of data that could not be practically coded by humans. In these settings, the researcher trains their own machine-learning model to replicate the human-classified labels with sufficiently accurate performance. For these approaches, researchers seek high out-of-sample precision and recall in retrieving human-classified labels. These approaches also have high levels of reproducibility because code can be provided. However, they still rely on human-classified labels, which are less reproducible and can have errors. Machine-coded classifiers can also have systematic biases and may not be generalizable to other datasets.

10.2.3 What This Method Does

Large language models generate text by predicting the next sequence of text from a text input. They can, therefore, be used to generate responses to questions such as, "On a scale of 1–10, 10 being most positive, how positive is the following text? 'It's a beautiful day!' Please provide only a number as output." In this way, researchers are using LLMs to label texts by providing specific instructions and codebooks to LLMs and evaluating many texts in sequence.

Large language models are "off-the-shelf" in that, unlike previous classifiers trained by the researcher, they are pretrained by large corporations or research groups. There are now a vast array of potential LLMs for social scientists to

PROOFS

consider using and many different options within each one. These choices will have impacts on the performance and cost of many annotation tasks. Our broad advice is that researchers empirically measure both the cost and the performance of LLMs on a task and pick the cheapest option that achieves what they consider to be an acceptable level of performance. We strongly recommend that researchers think carefully about what performance gains are worth significant cost increases. While more expensive models and settings often perform better than cheaper ones, this is not universal and should be tested for a scientist's exact use case. We talk more about validation steps later in this chapter.

To give an idea of just how much cost difference there can be, consider a scenario where the task is to label sentiment on social media platforms (aka "tweets") using a long codebook that takes up 5,000 tokens and the typical tweet is just 50 tokens.

- Option A: GPT-5.2 with high reasoning and one tweet passed at a time and no attempt to use batch processing or prompt caching.
- Option B: GPT-5-nano with no reasoning and 100 tweets passed at a time using the batch processing Application Programming Interface (API).

For Option A, a researcher would likely spend around $149,000 to code one million tweets. For Option B, they would spend $2.90–50,000 times cheaper (or, to put this another way, one could consider coding 50 billion tweets instead of 1 million for the same cost). It may be that Option B does not produce acceptable performance levels for a research question, but, with hundreds of thousands of dollars on the line, it is probably worth checking.[1]

The first choice is simply which provider and/or API to use. At time of writing, Open AI, Google, and Anthropic are close in price and performance on most benchmarks. There are also open-weight models to consider. These have the advantage of allowing full reproducibility because the exact model artifact can be stored, whereas closed-source providers may deprecate models or withdraw access to older versions over time (Barrie, Palmer, and Spirling 2025).

Open-weight models can be run on dedicated cloud providers such as replicate.com and openrouter.com. This makes their use largely similar to the closed weight providers except that it is usually possible to record the specifics of the analysis in enough detail that a result should be fully reproducible. However, the downside of this is that cloud providers of open-weight models typically offer little cost savings over using a closed weight model (while usually falling short on performance at a given price point). Another option for open-weight models is for researchers to run them on hardware they control. For smaller models this can be possible even on a standard laptop (Apple's recent

[1] The cost differences here come from a 25× difference in token cost, a 50 percent flex discount, saving nearly 99 percent of input token costs through input batching, and a 10× savings from avoiding the use of reasoning tokens.

PROOFS

Macbook computers are well-suited to this due to the high available memory for GPU processing). The setup for this is still relatively straightforward and can be managed through software such as Ollama and LM studio. For larger models, users typically need access to substantial computing infrastructure. These types of computing resources are not available across all universities, limiting this option to significantly fewer departments in practice.

The LLM landscape is changing fast, so the various cost/performance/reproducibility trade-offs will likely continue to change.

After deciding on a model to use, how should a researcher use it? While many people will be most familiar with LLMs through chat interfaces such as ChatGPT, we strongly recommend against using these interfaces for research purposes. Chat interfaces are opaque in what model is being called with what settings, are not well suited for large-scale tasks, and create room for copy-paste errors. Chat interfaces also create personalization, which means that the output can be specific to the history of a person's queries, which is not good for generalizability. We suggest that researchers use the APIs of the models for efficiency.

After choosing a model, the most important task is developing an effective prompt to ask the LLM to complete the labeling. In our experience, many initial LLM (and RA) failures result from having poorly described the task rather than the model being incapable of completing it. While early guidance suggested various tactics from telling the model to adopt a persona to threatening the model, the consensus has generally shifted toward clarity of the task, giving examples and iteratively improving the prompt. Large language models can be particularly susceptible to prompts that lead the LLM to produce a desired result, therefore, researchers should pay particular attention to the nature of the prompt (Anagnostidis and Bulian 2024).

Outside of the prompt, most LLM APIs will give several parameters that can be set that governs the output of the model. While a full description of all settings would be long and quickly out of date, two parameters are important to pay attention to: temperature and reasoning.

At the final layer of an LLM neural network, the model produces a score for each possible next token. These scores are translated into probabilities of choosing a token using a softmax function. This function can be tuned with a temperature parameter to spread the probabilities more evenly across potential next tokens (high temperature) or concentrate the probabilities on fewer tokens (low temperature). At the extreme, setting temperature to zero will typically make the output of an LLM deterministic by always picking the highest probability next token (though there are some nuances that mean this is not quite true in practice). In general, users should set the temperature to zero if the goal is to get close to reproducible labels. There are some circumstances where a scientist may want to choose a higher temperature if the variation across possible LLM responses is of interest (for instance, using multiple generations to find a consensus label).

PROOFS

Reasoning is a newer parameter. Traditionally, LLMs have simply generated the next tokens based on the input text fed through the neural network. However, AI researchers soon realized that LLMs could achieve better results for certain problems when told to "think step by step" (Wei et al. 2022). This has now been integrated into LLM architecture through an internal reasoning trace where LLMs first generate tokens that represent "reasoning" about a problem and then generate a final answer based on the input text and the reasoning steps that they wrote out.

An important factor for researchers to consider is that reasoning tokens are charged at the output token rate, which is typically around ten times the cost of input tokens. This means that the cost of reasoning can easily exceed the cost of the input text. Most models that support reasoning allow some level of control of the reasoning length either through a reasoning effort parameter (e.g., Open AI) or a reasoning token budget (e.g., Google). As usual, we recommend choosing the lowest level that achieves the desired performance.

One further choice that researchers should make is how many cases they pass to the LLM each time. Most people simply pass a set of instructions followed by a single case to label. However, this is cost inefficient because providers charge for passing the whole fixed instruction set (which can often be thousands of words) every time.

Some providers offer substantial discounts for repeated prompt prefixes, and asynchronous batch processing can also materially reduce costs. Another option is to pass more than one case at a time (input batching) (Cheng et al. 2023). For instance, Mellon et al. (2024) pass twenty-five open text responses at a time to the model. This spreads the cost of a long set of instructions over a large set of inputs. Importantly, using input batching can be combined with flex or batch processing discounts from providers that typically offer a 50 percent discount in exchange for a delay in processing the request. For many real-world tasks, input batching can achieve cost savings of greater than 90 percent. As always, it is important to validate performance when using these methods. Regardless of what choices a researcher makes, it is vital to document all decisions. Importantly, this includes choices that have default values. Researchers must record all parameters because it is not guaranteed that it will be possible to reconstruct what these were after the fact.

While much research in this area has used LLMs "off-the-shelf," many models have the option of inserting data that the researcher has labeled themselves to "fine-tune" the model. While fine-tuning requires significant human effort to create the training data, it can have substantial performance improvements for some tasks (Wang 2024; Waight et al. 2025).

Another important consideration when labeling data is data handling. Norms and interpretations of laws are not solidified in this area but, at a minimum, researchers should be careful that they do not put personally identifiable information into systems that will use that data for training. There are verified examples of LLMs reproducing training data in outputs (Kiyomaru et al.

PROOFS

2024), so a conservative assumption is that any data used to train an LLM has a small probability of being reproduced in its entirety in future. Large language model terms and conditions are complex and subject to change, so we recommend getting advice from within your institution.

While the most common uses of LLMs is for labeling text data, recent models can ingest other forms of information such as images, audio, or video. For example, this is an area that may be particularly fruitful for campaign studies, where LLMs could, for instance, classify advertisements with visual elements at scale. The best practices for labeling this kind of information are largely similar to labeling text, although we do recommend being mindful of the costs of using this kind of media, which can be high in comparison to simple text inputs.

Given the wide range of choices, researchers face when using LLMs, we strongly recommend testing a task using multiple models, settings, and parameters to find an option that represents a good balance between cost and performance.

10.2.4 Example Applications

Recent work illustrates both the promise and the range of LLM-based annotation in political science. Ornstein and colleagues (2025) provide one of the clearest general demonstrations. They revisit several established political text-analysis tasks and show that few-shot prompting can often substitute for more labor-intensive human labeling or conventional supervised pipelines, while requiring much less labeled training data and dramatically lower cost. One especially strong validation example comes from Wu, Nagler, Tucker, and Messing (2023), who use ChatGPT-3.5 to estimate the latent positions of US senators using pairwise comparisons and Bradley–Terry scaling. In that setting, they report very strong agreement with existing ideology benchmarks, suggesting that LLMs can recover meaningful latent structure when the objects being coded are highly salient and richly represented in public text.

At the same time, the strongest results so far often come from relatively favorable validation settings, and that should make researchers cautious about over-generalizing. The Wu et al. (2023a). application, for example, focuses on contemporary US senators – a high-profile set of political actors for whom both background information and external validation measures are unusually abundant. Likewise, Heseltine and Clemm von Hohenberg (2024) show that GPT-4 can perform very well when coding political text across multiple countries and tasks, comparing GPT-4 against human expert coding of tweets and news articles in the United States, Chile, Germany, and Italy. They report especially high accuracy on shorter texts such as tweets, with weaker performance on longer news articles and slightly lower performance outside English; they also show that a hybrid workflow, in which humans adjudicate disagreements across multiple GPT-4 runs, can improve reliability further.

PROOFS

These are impressive findings, but they also suggest a boundary condition: current evidence is strongest for relatively visible domains, shorter texts, and tasks where good benchmarks already exist. That means researchers should be careful in extending these successes to lower-salience politicians, less-covered institutions, lower-resource languages, historical corpora, or more weakly structured classification problems, where training-data coverage and independent validation may be much thinner.

Information retrieval tasks represent an adjacent field in the use of LLMs for annotation or data generation tasks. Mellon et al. (2024) show that language models can be used to code open-ended survey responses about the "most important issue" into substantive categories at scale, extending LLM-based annotation beyond elite political text and into survey research. Bol and Bono (2025), meanwhile, show that ChatGPT can estimate party positions in France with reasonable success when compared against contemporaneous expert judgments, particularly on major and salient dimensions. Taken together, these studies make a strong case that LLM-based annotation is already a useful methodological tool, especially when researchers need scalable coding at low cost. But they also imply that the next step is not simply to accumulate more success stories in easy cases. The more important challenge is to determine how well these methods generalize to harder settings – where labels are noisier, texts are longer or less standardized, actors are less prominent, and gold-standard validation data are limited.

10.2.5 Validation and Inference

Validation is critical when using LLMs for annotation because performance varies between datasets and tasks. Large language model outputs should be compared to human-coded data as validation that it is, indeed, measuring the concept as interpreted by humans. Like other machine learning workflows, performance should be evaluated "out of sample," that is, not using examples that have already been provided to the LLM or used in fine-tuning.

Many studies simply use agreement rates between LLMs and humans as the main metric to assess coder performance. However, this can be problematic for categories of rare events (e.g., hate speech in tweets) as a model can perform well simply by always choosing the largest category.

We recommend the use of balanced outcome metrics such as F1 (the harmonic mean of recall and precision) to reflect both types of performance. Other options include directly examining the confusion matrix.

One additional complication is that types of error matter more for different applications. False positives are especially problematic when the goal of the method is to characterize the population of positives in a low prevalence context. For instance, suppose that an LLM falsely classifies a tweet as hate speech 2 percent of the time but that the prevalence of hate speech tweets is only half of 1 percent.

PROOFS

Researchers should also consider that LLM errors in annotation may not be uniform throughout the data. In other words, the LLM may have more error for particular subsets of the data than others, which can cause bias in downstream analysis that uses LLMs as an input. We recommend using a de-biasing method to correct for those errors in downstream analysis (see Egami et al. 2023, 2024).

Because of the challenges of pure reproducibility when using closed-source models, we recommend that researchers take the following steps during validation:

First, researchers should always specify the exact model version (e.g., GPT-5.2.-2025-12-11 not GPT-5.2) as performance can vary meaningfully across releases from the same model family (Barrie, Palmer, and Spirling 2025).

Second, researchers should always revalidate when running new data collection even with similar models. To aid in this, we recommend that researchers set up their workflows to make swapping out model versions straightforward.

Third, as we discuss in the next section, researchers should make their workflows as auditable as possible.

Fourth, researchers should calculate and consider both cost and performance when deciding on models.

A challenge that researchers are beginning to face is what to do when models perform better than humans. Traditionally, we have validated machine performance against a benchmark of humans but, in some use cases, LLMs appear to outperform humans. Briggs et al. (2025) use a reconciliation approach where human RAs and multiple LLMs label the same content. Disagreements between coders of any type are then escalated to an expert human who reconciles the correct answer (while blinded to which answer was given by humans versus LLMs). A dataset of consensus codes and reconciliations is then used as the gold standard.

The situation we have not yet faced is when the LLM might have superior performance even than the reconciler. At that point, it is not clear how LLM performance could be validated. However, it would be likely that the reconciler approach would still indicate that the LLM was outperforming humans, but it would become difficult to tell by how much. We return to this point in Section 10.9.

10.3 LLM-GENERATED TREATMENTS AND STIMULI

10.3.1 The Method

Generating treatments and stimuli for an experiment is often the most unique and customized phase of experimental design and analysis. This is particularly true for designs using complex stimuli like text, images, or audio, where

PROOFS

the causal variable of interest is embedded within media that often simultaneously vary on many other dimensions. Researchers must think carefully about every aspect of what they show participants: how treatments differ from one another, what participants might infer beyond the intended manipulation, and what associations a given stimulus might evoke. The more complex the stimuli design – be it using videos with many aspects that could change between two treatments or a dynamic interaction – the more difficult it is to ensure that the variable researchers intend to test is, in fact, responsible for observed differences across conditions.

In practice, producing these treatments has traditionally required researchers to create or source all stimuli themselves. While best practices exist for various designs (Aguinis and Bradley 2014; Butler and Crabtree 2021), answering a specific research question still requires considerable discretion, and piloting and refining treatments can be slow and expensive. Some stimuli pose particular logistical challenges: creating images or videos where only a single element varies between conditions may be difficult or impossible; and training confederates for interactive experiments is costly, time-consuming, and difficult to precisely control.

To address these challenges, researchers have begun using LLMs for different forms and phases of treatment design. These include generating entirely novel stimuli (Evsyukova et al. 2025), creating new treatments based on existing examples (Porter and Velez 2022; Goldstein et al. 2024), tailoring treatments to individual respondents (Velez and Liu 2025; Hackenburg and Margetts 2024), and deploying an LLM as a conversational agent that directly interacts with participants (Argyle et al. 2023a; Yin et al. 2024; Argyle et al. 2025b). There is significant potential across these use cases – especially as relates to challenges of scaling treatments and having precise control over differences between treatments – that are difficult for the average researcher to generate, such as images. Yet they also raise new difficulties: validating construct fidelity at scale, navigating the ethics of LLM-participant interactions, and understanding whether automated generation itself affects treatment effects. Large language models expand what researchers can feasibly study, but at the same time they compound many existing challenges of experimental design.

10.3.2 Where We Came From

Generating experimental treatments has often been a creative exercise for researchers. This is particularly true for multidimensional treatments: Text-based treatments often vary in tone, length, and word choice; images commonly have different backgrounds and subject features; and treatments involving interactions with confederates commonly vary in content and style. Designs featuring these sorts of treatments confront a common challenge: isolating the causal dimension(s) of interest while holding others constant in

PROOFS

expectation. While the implications of this challenge are well-studied (e.g., Dafoe et al. 2018; Damann et al. 2025), the construction of treatments still involves considerable researcher discretion.

Experiments using multidimensional treatments have been particularly productive in the study of media and political communication, where multidimensional treatments are often necessary to effectively represent the media of interest. Text-based treatments, developed in survey experiments, enabled researchers to make strides in substantive areas involving shortform text (e.g., names in audit studies) as well as longer-form text (e.g., news articles). However, these text-based manipulations often change more than the causal variable of interest. For example, in an audit study, changing an applicant's name might also manipulate the reader's assumptions about race, ethnicity, and socioeconomic class (Landgrave and Weller 2022). Similar concerns arise with treatments involving human confederates, rule-based chatbots or bots on social media, and visual stimuli.

Large language models offer many potential benefits to these designs. Two especially promising possibilities emerge. First, by reducing the costs of text generation, LLMs make it possible to generate many more variants of each permutation of manipulated conditions, making it practical to use many treatments per condition rather than relying on a single carefully crafted exemplar. Second, in some cases, they improve the fidelity of the intended manipulation, particularly in designs where treatment is difficult to control, as with human confederates (Argyle et al. 2023a). But at the same time, at minimum, they simply transform the concern about multidimensionality and potentially amplify it. Researchers must now contend with learned associations in model outputs, correlated features of training data that a researcher often cannot observe, and the difficulties associated with auditing a larger volume of stimuli (Bai et al. 2025a). Nonetheless, the potential benefits have motivated rapid adoption, and Section 10.3.3 examines how researchers have applied LLMs to treatment generation.

10.3.3 What This Method Does

In practice, the basic operation is straightforward: Researchers provide instructions to language models specifying the content of stimuli, often explicitly attempting to hold some features constant and specifying which are meant to vary. However, this appeal rests on two mechanisms. First, LLMs can generate many instances of treatment. In practice, this may make it tenable to assume that idiosyncratic variations in nontarget dimensions of experimental stimuli average out. Second, a language model allows the researcher to instruct the model to explicitly hold certain features constant while varying the treatment along other dimensions. Yet the fundamental problem remains: instructing a model to vary one dimension does not guarantee other dimensions stay constant.

PROOFS

With this concern in mind, what is distinctive about LLM-generated experimental stimuli? Large language models transform the challenge posed by multidimensional treatments but do not eliminate it completely. The concern shifts from manually producing treatments varying only in the target dimension to ensuring that the model's learned associations do not introduce unintended covariation between treatment and potential confounders. Because models are trained on naturalistic data, prompts that are otherwise equal except for the target dimension may not yield outputs that only vary in the target dimension. As the subsequent examples demonstrate, it is still incumbent on researchers to validate experimental stimuli, though these examples also demonstrate the promise and potential benefits of LLMs in these use cases.

10.3.4 Example Applications

10.3.4.1 Generating Static Treatments

The most straightforward application is generating static treatments for randomized experiments. For example, consider audit studies: Scholars have critiqued the traditional approach of signaling race through names, noting that names may simultaneously convey information about ethnicity, socioeconomic status, and other attributes (e.g., Landgrave and Weller (2022), as discussed earlier). Evsyukova et al. (2025) sidestep name-based considerations by creating fictitious LinkedIn profiles with AI-generated images that vary applicant race directly. Profiles send connection requests, and researchers observe acceptance rates. The authors argue this approach addresses privacy concerns associated with using real photos, and crowdworkers could not distinguish generated images from real ones. Beyond privacy, AI-generated images could allow researchers to hold constant other features (e.g., background, clothing, age) and isolate the effect of race more precisely. However, the concern about learned associations applies here as well: Researchers must verify that generated images vary only on the target dimension, since models may introduce race-correlated features (e.g., background settings, attire, or image quality) that the researcher did not intend.

Beyond images, researchers have used LLMs to generate text-based treatments across a variety of research designs. For example, Goldstein et al. (2024) create many novel propaganda items based on existing examples, producing a volume that would be difficult to generate manually while allowing idiosyncratic variation to average out across items. Lai et al. (2026) conduct a conjoint experiment in which LLM-generated social media posts systematically vary anti-elite rhetoric, people-centric appeals, policy topics, and ideological direction. Porter and Velez (2022) use LLMs to create placebo control news stories that function as counterfactuals for articles selected to serve as treatment conditions. Velez and Liu (2025) use an LLM to scale treatments in a different way, generating counter-attitudinal information for each participant

PROOFS

in response to previous answers they had given in the survey. Similarly, Hackenburg and Margetts (2024) create persuasive messages micro-targeted to participants based on their demographic and political attributes. These applications enable research designs that would previously have been impractical, but they also raise validation concerns.

10.3.4.2 Interacting with Participants

Another promising direction is using LLMs as conversational agents to test theories that rely on interaction. Training human confederates for conversational experiments requires substantial resources and limits sample sizes; LLMs can interact with thousands of participants following consistent instructions. Argyle et al. (2025b) test how conversation length and message targeting affect political persuasion, instructing GPT-4o on the number of exchanges and whether to tailor arguments to individual conversation partners. They find no significant difference between messaging strategies, with statistical power that would be difficult to achieve without an LLM.

In addition to having persuasive conversations, some researchers have tested the abilities of language models to mimic emotional processes in conversation. Yin et al. (2024), for example, test whether conversations with an LLM could make people "feel heard." Other studies use LLMs as an assistant in person-to-person conversations, rather than as a substitute – for example, to moderate extreme language (Argyle et al. 2023a). These applications lower barriers to studying conversational processes, particularly for resource-constrained researchers, but raise concerns about validity, replicability, and ethics.

10.3.4.3 Other Uses

Though less central, there are several other ways in which researchers have begun to use LLMs to facilitate the experiment design process. In addition to generating treatments, some researchers have used LLMs to generate questions for use in experiments. Velez and Liu (2025) do so within an experimental context, developing specific questions for each participant to assess their reactions to the information which was generated for them based on their preexisting attitudes. Alternately, Götz et al. (2024) do so as a preparatory step, using GPT-2 to generate many potential scale questions, based on several starting questions, to measure a particular construct, which can then be validated as with any question battery as part of scale creation.

Finally, some researchers have proposed using LLMs not to generate experimental materials but to validate them. Ashokkumar et al. (2026) develop tools that estimate treatment effects and statistical power from proposed treatments and survey questions, which would typically require pilot studies or guesswork. However, such estimation likely works best for treatments and questions similar to those used in prior research. Estimates about novel stimuli that fall outside the model's training data would likely be inaccurate.

PROOFS

10.3.5 Validation and Inference

Though LLMs offer unique opportunities for generating experimental materials, their use compounds existing challenges associated with writing stimuli. The first is construct validity. When designing treatments manually, researchers must ensure that the intended causal dimension is what changes and that participants observe only this manipulation. Validating LLM-generated stimuli follows a similar logic but is complicated by the very features that make LLMs attractive: scale and complexity. When generating many treatments, at what point can researchers be confident that successive outputs accurately represent the target construct? When LLMs interact with participants, how many input permutations must be tested before researchers trust that all participants receive valid treatment?

A related challenge is minimizing unintended covariation – what is sometimes called "bundled" treatments – where multiple dimensions vary simultaneously and their effects cannot be separated. Isolating the target dimension while holding nontarget dimensions constant is difficult with any multidimensional treatment; LLMs introduce two additional complications. First, as noted previously, validation becomes harder at scale. Ensuring that treatments vary only on the target dimension is simpler for single sentences than for paragraphs, conversations, images, or videos, yet LLMs make generating these more complex stimuli easier. Second, as with many phases of research design, the process of designing treatments or training human confederates often forces the researcher to clarify theoretical expectations. Crafting treatments by hand forces researchers to consider which elements matter, what mechanisms might operate, and what unintended cues participants might detect. Large language model generation may shortcut this process, potentially to the study's detriment. Finally, LLM-generated treatments introduce a novel concern: disclosure effects.

Several studies show that when participants know they are interacting with an LLM, treatment effects are attenuated (Yin et al. 2024; Palmer and Spirling 2023). This raises questions about whether findings generalize to non-LLM contexts. Additionally, unmonitored LLM interactions with participants pose ethical concerns, particularly for longer or more sensitive conversations.

Given these challenges, extensive validation must be central to any use of LLMs for treatment generation. When generating a relatively small number of stimuli, researchers should identify the exact elements that they believe are responsible for any observed effects and justify other variation between elements. For larger scale treatment generation, all LLM output should be recorded and released as a component of the final data. Such output should itself be analyzed as media to assess feature variance across treatments, variance within treatments, and how this correlates with final outcomes. Finally, where possible, unless LLMs are the subject of the study, LLM-generated treatments should be compared to human-generated treatments, even if on a relatively small scale.

PROOFS

10.4 SILICON SAMPLING

10.4.1 The Method

Silicon sampling, also known as digital twins (Toubia et al. 2025; Peng et al. 2025), is a research methodology in which LLMs are used to generate synthetic proxies or replicas of individual-level research data (Argyle et al. 2023b; Kozlowski and Evans 2025; Chapter 3). The basic approach involves prompting (or otherwise training or tuning) a language model to provide a response that approximates some unknown quantity on the basis of provided information that constructs a particular "persona." For example, the model might be provided with party identification, race, gender, and other demographic or attitudinal characteristics and asked to provide the expected vote choice for that constructed individual. The provided persona constructions vary in complexity, and the generated data can be used to proxy specific questions, full surveys, semi-structured interviews, or responses to experimental stimuli. The purpose is to generate lower cost alternatives to inform, augment, or replace human subject data collection in behavioral science research.

Silicon sampling is controversial on both practical and theoretical grounds. On the practical front, there is significant debate about the extent to which political and social phenomena can be fully and completely represented by LLMs. Existing research provides demonstrations that are both more optimistic (Argyle et al. 2023b; Dillion et al. 2023; Horton 2023; Goli and Singh 2024; Ashokkumar et al. 2026) and more pessimistic (Peng et al. 2025; Santurkar et al. 2023; Boelaert et al. 2025; Bisbee et al. 2024; Wang et al. 2025a; Qu and Wang 2024).[2]

However, proponents cite the lower cost (Broska et al. 2025), instantaneous data, and potential for exploration of questions that may be challenging or unethical to explore with human subjects. This is even more poignant as widespread LLM use by online panelists leads to questions about data quality in existing sources and the likelihood of rising costs for human survey data in the future (Westwood 2025; Veselovsky et al. 2025). Both start-up companies and established survey panel providers are promoting products for marketing insights using completely in silico data.

On theoretical grounds, there is also an important concern about the validity of replacing humans in research on human attitudes, psychology, and behaviors. This reflects concerns about how the fundamental nature of the scientific study of people and communities is distorted when researchers no longer directly observe or interact with their human subjects. Additionally, even if scholars can demonstrate that silicon sampling methodologies provide good replications in a particular application, it remains unclear what standards

[2] See also Section 10.3.4.

PROOFS

would provide enough confidence in synthetic data to use it for novel scientific inference without validation against a human sample.[3]

10.4.2 Where We Came From

Different groups of scholars bring different intellectual traditions to their assumptions and approaches regarding silicon sampling. There are at least three major traditions in this work.

First, some authors approach silicon sampling from a perspective of deriving insights from second-hand uses of "found" big data (see Grimmer 2015). Second, other authors approach the data from a tradition of unrepresentative samples and online crowd work (Porter et al. 2020; Kittur et al. 2013). Finally, still other scholars bring a lens of agent-based modeling. The first two traditions are briefly reviewed here. Readers are directed to Section 10.5 for a more complete discussion of agent-based modeling and LLM simulation.

Large language models are trained on massive swaths of written human text (and multi-modal audio or visual inputs for some models). Thus, some scholars approach LLM simulation from the conceptual background of machine-learning algorithms and "big data." Because human authors produced (at least a nontrivial amount of) the training material and because the LLM learns many patterns of meaning and conceptual association that go far beyond just rote encoding of rules of grammar (Wei et al. 2022), some of the associations encoded into the LLM weights reflect human patterns of thinking, behaving, speaking, and valuing (Kozlowski and Evans 2025). Unwanted and undesirable emergent patterns are termed "algorithmic bias" (Bolukbasi et al. 2015). Large language model–generated content reproduces known patterns of partisan, racial, gender, national, or other biases (Gallegos et al. 2024; Chapter 5; Chapter 6). However, while normatively undesirable in society, valuable social scientific patterns, termed "algorithmic fidelity" by Argyle et al. (2023b), may be encoded alongside the biased expressions of human psychology and behavior. Scholars in this approach may view interacting with the LLM as an exercise in navigating the black box to effectively surface the right associations from the trillions of potential parameters in a model. Thus, like many big data applications, although the whole body of an LLM's output may be biased and unrepresentative (Santurkar et al. 2023; Wang et al. 2025a), scholars may be able to carefully construct prompts and interactions that elicit scientifically meaningful representations of populations and behavioral processes.

Another set of scholars comes to silicon sampling from a perspective of crowdwork and adapting to available but unrepresentative – and sometimes unreliable – respondent populations. Large language models are agents trained to do particular tasks, but they may not naively be representative of whole populations. Prompts, fine-tuning, and other methods of instructing the LLM

[3] See Chapter 3 for much more in this regard.

PROOFS

provide guidelines for LLM behavior which, when coupled with their massive training backgrounds and emergent abilities, enable them to engage in a wide range of tasks including those related to silicon sampling. While not specifically conducting silicon sampling, Vafa et al. (2025) conceptualize the importance of LLM steerability – meaning that models provide appropriately varied output based on different prompts – and propose a novel metric for evaluating model steerability. If models are appropriately steerable, then they can provide a range of different attitudes that may be sufficient to reflect a whole population. However, it is not clear what the sampling frame is for an LLM or how to think about randomness or representativeness in sampling from that black box. Authors in this space view statistical calibration of results as a necessary, but likely not sufficient, research step to improve the validity and reliability of output (Vafa et al. 2025; Hullman et al. 2025; Egami et al. 2023; Imai and Nakamura 2025; Broska et al. 2025).

10.4.3 What This Method Does

Silicon sampling is modeled on survey research, and the end result is typically an expected individual-level dataset but with a radically different data-generating process. At their core, LLMs are next-word prediction engines. Given a prompt with certain characteristics, the LLM is expected to respond to further survey questions in a way that is likely, given the "persona" it received as an input. Because different words in a prompt produce different probabilities for the next token, a single LLM can be used to produce a wide variety of outputs that generate expected patterns of variation based on the prompts.

The most basic approach to silicon sampling sets up a template for a persona (often using a "system" prompt to create an overarching set of instructions), which pipes in standard fields for a variety of descriptive characteristics of a given persona and asks the LLM to produce an additional piece of information. In some cases, the personas represent a complete set of possible profiles given the possible intersections of the various background information, with each persona-prompt being run multiple times to create a distribution of responses. In other cases, researchers benchmark an existing survey data source, using actual demographic profiles of respondents to surveys, such as the American National Election Study or World Values Survey, to generate persona prompts that are "digital twins" of human subjects.

Although this sounds straightforward, there are several essential decisions that can have important implications for the survey results. While these are by no means the only relevant decisions, two crucial researcher choices are: the information contained in the prompt and model selection.

First, what information should go into the prompt? When human subjects answer survey questions, survey researchers assume the subject exists as a unique and independent entity and the purpose of survey questions is to objectively observe and measure various aspects of that person as they really exist.

PROOFS

There are well-known exceptions to this assumption in survey research, such as a "Hawthorne effect" in which the mere act of being observed changes a behavior. Social desirability bias, interviewer effects, and other survey design and question wording decisions affect whether it is possible to observe the survey taker without influencing the construction of the responses. But most survey research practices aim to minimize the impact of these researcher interferences (Weisberg 2009).

In silicon sampling, however, the persona is a digital representation of a real or hypothetical person who is constructed entirely through the training and prompting of the model. Thus, information that is not directly given to the model in the prompt cannot be assumed to be known or a part of that person's identity, in the same way that questions not asked by the researcher can still be assumed to be constitutive of the human. For example, the gender of a human survey respondent would be assumed to have an impact on their worldview even in a survey that does not record the respondent's gender. But if gender is not provided in the prompt to an LLM, the persona does not have a defined gender and the LLM will use an unknowable process of probability calculation to account for that missing information (see discussion in Gui and Toubia 2023). That is all to say: giving additional and more relevant information to the LLM is likely going to improve prediction performance when compared with giving less information. However, it might also lead to the model placing outsized importance on enumerated background characteristics or result in the information provided being so comprehensive as to render the model contribution trite.

10.4.4 Example Applications

There are two important ways that researchers have pushed the methodology forward in giving more information to the LLM for the purposes of silicon sampling. Some authors provide backstory information that is substantially more detailed (Park et al. 2024; Toubia et al. 2025). Park et al. (2024) conducted multi-hour in depth interviews with 1,000 respondents to incorporate a rich history into silicon sampling generations that is not typically available in a survey. Using this information in the prompts resulted in high performance on prediction of later survey questions.

Another approach to providing more information to the model is moving beyond using only the prompt to generate the persona and, instead, directly fine-tuning the model (Binz et al. 2025; Kim and Lee 2023; Suh et al. 2025). Fine-tuning is a process of using an end-stage training data set to adjust the model weights to incentivize behavior that more closely matches a specific, desired outcome set for a particular task. Fine-tuning requires higher technological capacity and computing resources, it may unexpectedly not only affect other model functions (Luo et al. 2025) but can improve performance enough to be worth the investment in some cases.

PROOFS

Second, in addition to determining the process and information, the model will use to construct the persona, the researcher must also select a model to use. One primary decision is whether to use an open-weight model or a closed-weight model (see more discussion of the trade-offs in Section 10.4.5). However, even beyond this high-level decision, there is evidence that differences in training across model families – and post-training steps within model families – can have important, direct effects on model performance for silicon sampling tasks (Lyman et al. 2025; Boelaert et al. 2025). For example, models of different sizes, in different families, or with different post-training may have different biases in the response distributions, different capacity to complete the task, or different refusal triggers. One important implication of this heterogeneity is that prompting should be done in a model-specific way, and researchers may need to craft different prompts to get comparable performance from different models, as different models may interpret an identically worded prompt to be suggesting a qualitatively different kind of task. This is consistent with research demonstrating that even relatively minor differences in prompt structure, such as wording or ordering of otherwise identical information, can have a substantial impact on silicon sampling output (Brucks and Toubia 2025).

10.4.5 Validation and Inference

Existing empirical demonstrations of silicon sampling come to many competing conclusions about the potential success, failure, and biases or limitations of the approach. In addition to variation that is introduced by model selection and prompting strategies, nascent research in this space has not yet coalesced around an appropriate set of metrics. Existing validation approaches include individual-level prediction accuracy metrics (F1 scores, percent matches, or correlations between human and silicon observations), distributional characteristics (means, proportions, or variances), or relational statistics (subgroup analyses, treatment effects, or coefficient estimates between independent and dependent variables). Kozlowski and Evans (2025) provide a useful framework of potential pitfalls and associated validation practices for silicon sampling.

Choices of different metrics for validating silicon sampling are signals of fundamentally different inferential goals across studies. Thus, determining which is the appropriate metric depends on clearly identifying the kind of inference that scholars want to make using the synthetic data (Argyle et al. 2025a; Chapter 3). For demonstration purposes, Sections 10.4.5.1–10.4.5.3 briefly describe three examples of different underlying assumptions about the inferential purpose of a silicon sampling application that imply different validation metrics.

10.4.5.1 Individual-Level Imputation

As noted in Section 10.3, Kim and Lee (2023) present a fine-tuning-based approach to silicon sampling for data imputation tasks. In an imputation task,

PROOFS

the silicon data will be used alongside human data to make statistical inferences about the attitudes or behaviors of specific survey respondents. Thus, the goal is to exactly predict the response that a specific given respondent would have had to the question at hand, had they provided it at a given point in time. In this case, the authors choose to use metrics that test prediction accuracy by directly comparing a specific human response to the probability that the model selected that response using an AUC (Area Under the receiver operating characteristic Curve). Other scientists use more standard metrics of proportion agreement or Cohen's kappa (Qu and Wang 2024).

10.4.5.2 Simulations of Aggregated Distributions

In other applications, the goal is not necessarily to predict a single individual precisely but rather to replicate the distribution of opinions in a given population or subpopulation. Many early attempts at silicon sampling compared measures of central tendency, such as proportions or means, from synthetic data to human data finding clear signals (Argyle et al. 2023b). However, later scholars emphasized different metrics of variation or distributional overlap as important factors for inferences based on distribution, finding that silicon samples often lacked the required variation. Specifically, Bisbee et al. (2024) note that, if researchers want to use silicon data for pretesting or power analysis, then variance is an essential element for valid inferences. Variation in responses can be dramatically affected by both temperature (which is subject to researcher control) and post-training alignment process (typically outside of researcher control) (Kirk et al. 2023), making model selection and evaluation essential when prioritizing distributional variation.

10.4.5.3 Simulations of Relationships

Still other researchers want to use silicon sampling methodology in a causal inference toolkit to investigate potential relationships or treatment effects (Binz et al. 2025; Imai and Nakamura 2025; Gui and Toubia 2023). From an observational perspective, one approach is to use subgroup analysis or estimated coefficients to carefully investigate whether silicon samples are able to produce expected relationships between key independent and dependent variables (Santurkar et al. 2023; Argyle et al. 2023b). Typically, that approach finds that silicon samples are more accurate for majority groups or Western countries (Wang et al. 2025a; Santurkar et al. 2023; Qu and Wang 2024). From an experimental perspective, scholars have evaluated whether silicon samples can reproduce treatment effect estimates (Gui and Toubia 2023; Aher et al. 2023; Ashokkumar et al. 2026; Horton 2023), perhaps with additional statistical corrections (Broska et al. 2025; Hullman et al. 2025). If such effects could be reliably estimated, then LLMs might provide useful tools for piloting experiments or selecting treatment materials. While many (but not all) silicon experiment replications produce expected directional effects, effect size can be incorrect (Cui et al. 2025). Furthermore, the presence of previously published

PROOFS

experiments in training data raises questions about the validity of the simulation, known as data leakage (but see Ashokkumar et al. 2026 who find similar performance simulating published and unpublished experiments).

Other researchers may bring additional motivations to silicon sampling approaches, which may imply still other appropriate qualitative evaluation or statistical tests. While best practices and community standards are still being developed, researchers are encouraged to be explicit about the inferential purpose of their silicon sampling approaches and to explicitly state and justify the selection of particular evaluation standards that support the desired inferences (Argyle et al. 2025a).

10.5 AGENT-BASED MODELS AND GENERATIVE AI

10.5.1 The Method

Agent-based models (ABMs) are computational simulations composed of multiple autonomous entities ("agents") that perceive their environment, maintain internal states, and select actions. The methodology proceeds by specifying (1) a population of heterogeneous agents; (2) interaction structures such as networks, spatial environments, or institutions; and (3) behavioral rules and update schedules. Researchers then observe the macro-level patterns that emerge from repeated local interactions (Bonabeau 2002; Macy and Willer 2002; de Marchi and Page 2014).

The core theoretical commitment is generative: ABMs seek to grow collective phenomena from plausible micro-foundations, rather than stipulating them and learning from the distribution of outcomes across many simulation runs (Epstein and Axtell 1996; Axelrod 1997). This contrasts with approaches that treat collective outcomes as simple aggregations of individual attributes, emphasizing instead how macro-level patterns arise from micro-level interactions.

Generative ABMs (GABMs) – sometimes called "agentic" or "multi-LLM" systems – replace hand-coded behavioral rules with LLMs. Instead of explicit heuristics or equations governing agent choice, an LLM serves as a context-sensitive policy function: Given a role, goals, memory, and institutional constraints, it generates the agent's next action (typically communicative) and updates its internal state (Park et al. 2023; Lu et al. 2024).

Crucially, this involves not just an LLM but an entire architecture: Prompts encode roles and constraints, memory modules compress experience, orchestrators manage turn-taking and tool access, and evaluators assess outputs (Park et al. 2023; Wu et al. 2023b; Zhou et al. 2024). This approach enables new forms of expressiveness: Language, persuasion, justification, and norm enforcement become endogenous simulation dynamics rather than exogenous labels or post-hoc interpretations. That is, political concepts – ideology, framing, persuasiveness – can be operationalized through generated linguistic behavior (and then measured), rather than only as fixed scalar attributes.

PROOFS

10.5.2 Where We Came From

Schelling's segregation model (Schelling 1971) remains the canonical example of agent-based modeling in social science because it demonstrates the power of interaction-driven emergence. Schelling showed that extreme residential segregation can arise even when individuals hold only mild preferences for same-group neighbors – a macro-pattern that cannot be predicted from simply aggregating individual tolerance levels. This demonstration of how simple micro-rules generate complex macro-outcomes catalyzed broader adoption of ABMs across the social sciences.

By the 1990s and early 2000s, ABMs had established a substantial lineage in political science, sociology, and economics (de Marchi and Page 2014; Macy and Willer 2002). More recently, this tradition has cross-fertilized with computational social science approaches that leverage large-scale digital trace data as input for formalized models of network dynamics and information flows (Lazer and Ognyanova 2024; Lazer et al. 2020).

Since the consolidation of ABMs as a generative methodology, a series of influential contributions demonstrated how simulation can link micro-level rules and interaction structures to macro-level regularities. Epstein and Axtell (1996) offered a paradigmatic demonstration of "growing" societal-level patterns from simple behavioral and environmental primitives, while Axelrod (1997) showed how adaptive agents engaged in repeated interaction can generate robust patterns of cooperation and norm-like behavior. Work in a similar spirit used computational models to account for specific empirical regularities: de Marchi (1999) formalized how adaptive dynamics and feedback can produce electoral instability, and Cederman (2003) connected micro-level interaction mechanisms to macro-level distributions in war severity.

As we will see further on, contemporary generative ABMs exhibit a comparable breadth of ambition, ranging from society-scale simulations of interacting populations to narrowly targeted models designed to probe why particular outcomes emerge under specified institutional and informational constraints.

Canonical political-economy ABMs specify four components: (1) a heterogeneous population of agents; (2) an environment (spatial, network, or institutional) governing interaction patterns; (3) micro-rules mapping local observations to actions – typically boundedly rational heuristics rather than full optimization; and (4) state-transition functions updating agents and environments after each interaction. Whether agents play iterated strategies in repeated games (Axelrod 1997) or follow simple behavioral rules on spatial grids (Epstein and Axtell 1996), the research logic is identical: characterize how macro-outcomes vary with micro-rules, interaction structures, and initial conditions across many simulation runs.

The traditional ABM approaches offered political scientists three advantages: interpretability by construction (explicit behavioral rules made mechanisms traceable), computable institutions (formal rules became manipulable

PROOFS

parameters for counterfactual analysis), and experimental control (systematic variation of micro-assumptions revealed their distributional consequences) (Epstein and Axtell 1996; Axelrod 1997; de Marchi and Page 2014).

These methods nonetheless struggled with behavioral realism: Compact decision rules could not capture linguistic and interpretive political behaviors like persuasion, framing, or norm contestation, which were reduced to scalar updates or thin heuristics. Measurement alignment also proved difficult: Internal states rarely mapped cleanly to observables, complicating validation. And because ABMs prioritize mechanistic explanation over prediction, external validity claims remained limited to "mechanism X can generate pattern Y" without strong forecasting power (de Marchi and Page 2014).

As we will see later, GABMs are able to overcome some of the problems with behavioral realism. Specifically, we can use these technologies to populate a system with agents that are endowed with the ability to interact with language rather than some pre-specified distributional formalism. In so doing, however, we also introduce problems relating to mechanistic interpretability and replicability. Overall, the "generative" shift with LLMs should be read as changing (1) the representation of agent cognition and discourse and (2) the practical workflow for building and iterating environments. It does not replace core ABM standards: Explicit mechanisms, controlled experiments, and transparent documentation remain the baseline.

10.5.3 What This Method Does

So, how are practitioners using generative AI for agent-based modeling? Pseudo-code 1 and 2 describe the higher-level attributes of the traditional ABM and more recent GABM approaches (see Figure 10.1). In the ABM panel, heterogeneous agent attributes parameterize explicit decision rules that map observations to actions, after which interactions generate outcomes and state transitions that can be aggregated across runs. In the GABM panel, personas and memory retrieval form part of the agent's context for a black-box LLM policy that emits both natural-language output and structured actions, with transcripts and logs becoming additional artifacts for measurement and replication.

This points to three components of new GABM approaches that distinguish these methods from prior agent-based modeling. First, instead of representing heterogeneity primarily as numeric attributes or fixed types, GABMs often define agents through personas: natural language profiles that encode identities, priors, goals, and situational constraints. Personas can be coarse (e.g., membership in the red team versus the blue team) or richly parameterized from survey microdata and other sources, allowing researchers to inject structured variation into populations while still enabling context-sensitive linguistic behavior (Park et al. 2023; Törnberg et al. 2023).

PROOFS

Algorithm 1 Toy ABM (hand-coded rules)

Inputs: agent attributes (θ_A, θ_B); initial states (x_A, x_B); environment S_0
for $t = 1$ to T **do**
  $o_A, o_B \leftarrow \text{OBSERVE}(S_{t-1}, x_A, x_B)$
  $a_A \leftarrow f_A(o_A; \theta_A)$; $a_B \leftarrow f_B(o_B; \theta_B)$
  $yt \leftarrow g(a_A, a_B, S_{t-1})$ ▷ micro outcome
  $(x_A, x_B, S_t) \leftarrow h(x_A, x_B, S_{t-1}, yt)$ ▷ state transition
end for
$\widehat{M} \leftarrow \text{AGGREGATE}(\{yt\}_{t=1}^{T})$ ▷ macro outcome

Algorithm 2 Toy GABM (LLM agents)

Inputs: *personas* (p_A, p_B); goals; initial states; environment S_0; memories M_A, M_B
for $t = 1$ to T **do**
  $m_A \leftarrow \text{RETRIEVE}(M_A, S_{t-1})$; $m_B \leftarrow \text{RETRIEVE}(M_B, S_{t-1})$
  $a_A, \text{text}_A \leftarrow \text{LLM}(p_A, o_A, m_A)$ ▷ black-box policy + language
  $a_B, \text{text}_B \leftarrow \text{LLM}(p_B, o_B, m_B)$
  $M_A \leftarrow M_A \cup \{S_{t-1}, \text{text}_A, a_A\}$; $M_B \leftarrow \cdots$
  $yt \leftarrow g(a_A, a_B, S_{t-1})$
  $(S_t) \leftarrow h(S_{t-1}, yt)$ ▷ state transition
end for
$\widehat{M} \leftarrow \text{AGGREGATE}(\{yt\})$ and / or $\text{MEASURE}(\{\text{text}_t\})$

FIGURE 10.1 Pseudo-code: traditional agent-based model and generative agent-based model architectures.

Second, many GABMs implement persistent memory as retrieval over prior experiences, observations, and interactions. Whereas classic ABMs typically store state as a small vector (e.g., beliefs, payoffs, or strategy parameters), generative agents store compressed natural language records of events and then selectively retrieve them when acting. This makes it easier to represent path dependence in communication and interpretation (e.g., reputations, grudges, learned narratives, agenda carryover), but it also introduces new modeling choices that must be specified and logged (what gets stored, how it is summarized, and what is retrieved at each step).

Third, GABMs replace the hand-coded decision rule with an LLM policy: a stochastic, black-box mapping from context to actions (and often to accompanying language). In practice, the agent's "policy" is implemented as a prompt bundle (system rules + persona + local observations + retrieved memory) passed to an LLM whose outputs depend on sampling and decoding settings (e.g., temperature, maximum tokens). Because unconstrained generation can drift or violate required formats, GABMs often add guardrails: constrained action schemas, validators/parsers, retry and fallback logic, and sometimes an adjudicator that enforces institutional rules. Methodologically, these decoding and control parameters are part of the model specification, since they shape variance, path dependence, and the stability of multi-step trajectories.

A fourth component, often implicit in traditional ABMs but explicit in GABMs, is the orchestration and evaluation components. Orchestration refers

PROOFS

to the controller that schedules interactions (who speaks/acts and when), gates observations (what an agent is allowed to see), assembles prompt context (personas, retrieved memory, institutional constraints), and routes tool access and guardrails. Evaluation refers to the measurement layer that audits agent behavior and converts language traces into analyzable outcomes (e.g., by classifying text into states or scoring toxicity/civility), typically alongside ablations and robustness checks. These components are visible in the applied examples: Park et al. (2023) rely on an environment server and planning/memory loops plus ablation-based assessment; Chuang et al. (2024) enforce a structured interaction schedule and map tweets into discrete opinion states via classification; and Törnberg et al. (2023) treat the feed-ranking rule as the orchestrated institutional mechanism and operationalize outcomes from generated text under controlled counterfactuals.

Together, these components change what ABMs can represent. Classic ABMs often reduced political interaction to scalar belief updates or stylized heuristics because rich interpretation and language were expensive to model explicitly. By contrast, GABMs make it comparatively cheap to model political behavior that is mediated by text: persuasion and counter-persuasion, framing contests, threat and reassurance signaling, norm contestation, and so on. The inferential logic remains ABM-like – macro-patterns are still functions of micro-assumptions and interaction structures. But the micro-foundations can be substantially richer in settings where meaning-making and communication are central and where researchers may wish to treat the evolving content of interaction as both a causal channel and an empirical object of measurement.

10.5.4 Example Applications

Emerging GABM applications can be organized by simulation scope: (1) open-ended "whole society" environments generating rich interaction traces; (2) macro-phenomena (polarization, consensus) as emergent outcomes of language-mediated network interaction; and (3) targeted institutional interventions comparing counterfactual rule designs.

10.5.4.1 Whole-Society Simulation

Park et al. (2023) populate an artificial town with twenty-five LLM agents, each maintaining memory streams, retrieving relevant experiences, and periodically synthesizing higher-level inferences. The environment constrains observation (locality) and interaction (turn-taking), while language traces (conversations, plans) are logged as analyzable artifacts. Their canonical demonstration – a party organized through conversational diffusion without scripted coordination – illustrates how social organization can emerge from transparent micro-rules. Methodologically, this establishes a recipe for open-ended emergence with inspectable instrumentation.

PROOFS

10.5.4.2 Macro-level Opinion Dynamics

Chuang et al. (2024) adapt classical opinion-dynamics ABMs to language-mediated interaction: Agents exchange tweets conditioned on personas, update beliefs through verbal responses, and produce discrete opinion states via classification. This design recovers familiar ABM metrics (trajectories, fragmentation) while retaining linguistic traces of framing and persuasion. Critically, they diagnose that default LLM behavior converges toward consensus unrealistically, then show how prompt engineering and memory design can reintroduce confirmation bias to match empirical polarization patterns. This provides a template for macro-phenomena GABMs with expressive agents and comparable metrics.

10.5.4.3 Institutional Counterfactuals

Törnberg et al. (2023) simulate Twitter-like platforms with agents initialized from survey microdata (American National Election Survey demographics, attitudes, usage). The intervention varies ranking algorithms (engagement-ranked within-network, global engagement, cross-partisan bridging), holding all else constant. Agents post, share, and engage at calibrated rates, with feedback loops between content and visibility. The GABM contribution is operationalizing outcomes directly from generated text (toxicity, cross-partisan engagement) while retaining full transcripts for qualitative audit – a computational sandbox for institutional design.

10.5.4.3.1 ADDITIONAL APPLICATIONS Related work includes society simulators scaled to thousands of agents for policy interventions (Piao et al. 2025; Wang et al. 2025b), election models simulating vote choice under campaign inputs (Zhou et al. 2025), and collective-action cascades with mixed LLM and simple agents (Mou et al. 2024). Reusable platforms are emerging to standardize environments and evaluation (Tang et al. 2025; Zhou et al. 2024; Wu et al. 2023b).

More critical work emphasizes the need for empirical benchmarking and transparency around proprietary models (Palmer et al. 2024; Bail 2024; Barrie and Törnberg 2025; Münker et al. 2025) and raises doubts about the ability of LLMs to faithfully recreate human-like discourse (Pagan et al. 2025). It is to these and related questions we will turn in the final two parts of this chapter section.

10.5.5 Validation and Inference

Generative ABMs inherit traditional ABM challenges – stochastic outcomes, path dependence, and specification sensitivity – while adding new ones: Micro-rules are partly black-box language models where small prompt or decoding changes can produce qualitatively different trajectories (de Marchi and Page 2014; Larooij and Törnberg 2026). Claims should therefore be distributional and conditional: about mechanisms and outcome patterns given an explicit

PROOFS

design (personas, memory, schedule, model version, sampling parameters), not about single-run narratives.

Following established ABM methodology (de Marchi and Page 2014; Windrum et al. 2007), we propose separating verification (does the system behave as specified?) from validation (does it capture the target phenomenon for its stated purpose?). We provided a bulleted list in Sections 10.5.5.1 and 10.5.5.2 of exactly what we mean by this.

10.5.5.1 Verification

Treat GABMs as experimental systems requiring robustness checks:

- *Sensitivity*: vary prompts, temperature/top-p, interaction order, topology, memory settings.
- *Ablations*: remove memory, reflection, tools, or roles to identify active components.
- *Cross-model robustness*: test whether core patterns persist across model families or versions.

10.5.5.2 Validation

When claiming realism or policy relevance, specify validation targets and calibration strategies:

- *Stylized* facts: match known regularities (fragmentation, homophily, engagement distributions).
- Empirical benchmarks: compare outputs to survey marginals, network metrics, or text corpora when representing specific populations (Larooij and Törnberg 2026).
- Contamination checks: test whether emergent patterns reflect training-data leakage rather than genuine emergence (Ashery et al. 2025; Barrie and Törnberg 2025).
- No circular evaluation: avoid LLMs grading their own outputs as primary validation (Larooij and Törnberg 2026).

10.6 TOOL USE AND WEB-BROWSING: LLMS AS RESEARCH AGENTS

10.6.1 The Method

Large language model–driven tool use describes a set of research workflows where LLMs are connected to external tools, such as web search, browsing, file readers, and code execution, to help researchers find, collect, and organize data. In this setup, the LLM is not primarily doing the final analysis. Instead, it acts as a flexible interface that takes a researcher's instructions and turns them into concrete steps: searching for sources, navigating websites, extracting relevant information, and assembling the results into structured datasets.

PROOFS

One major advantage of this approach is that it makes it much easier to achieve labor-intensive tasks, such as compiling large datasets, that would normally require a great deal of manual work. For example, a task like collecting biographical information for every member of a legislature, such as the university they attended or their town of birth, would traditionally involve many hours of web searches, copy-pasting, and error checking. With LLM-driven tool use, the same task can often be completed in minutes by combining search, browsing, and extraction tools in a single workflow (Mao et al. 2023; Nakano et al. 2021). This substantially lowers the fixed cost of general information collection and the building of custom datasets, as well as making it easier to iterate and adjust as research questions evolve.

Another important feature of this method is its ability to work with non-text or semi-structured sources. Contemporary LLM systems can extract information from PDFs, scanned documents, tables embedded in images, and other formats that are difficult to handle with traditional scripts (Xu et al. 2020; Kim et al. 2021; Huang et al. 2022; OpenAI 2023). For social scientists, this opens up new kinds of data, such as archival records or administrative documents that were previously too time-consuming to process at scale. In addition, LLMs increasingly function as coding assistants, helping researchers write and adapt scripts for scraping websites or collecting data automatically (Ahluwalia and Wani 2024). This reduces both the time and technical expertise required to build research pipelines, making these methods accessible to a wider range of scholars.

10.6.2 Where We Came From

Historically, the kinds of tasks now handled by LLM-driven tool use were carried out almost entirely by hand. Data collection in the social sciences typically involved researchers or hired research assistants searching for information online (or in person), downloading (or xeroxing) documents, and manually entering values into spreadsheets or databases. Projects that required assembling attributes across hundreds or thousands of cases often took weeks or months and required careful supervision to ensure consistency. As a result, due to temporal or financial constraints, the scale of data collection frequently constrained the kinds of research questions scholars were willing to pursue.

Attempts to automate these workflows existed, but they usually depended on custom scripts written for very specific tasks. Web scrapers or document parsers could speed things up, but they required substantial programming expertise and were often fragile. Small changes to a website's structure or unexpected variation in document formats could cause an entire pipeline to fail. This meant that automation was unevenly adopted and often concentrated among researchers with strong technical backgrounds or access to specialized collaborators.

PROOFS

Both manual and scripted approaches also raised concerns about accuracy and reproducibility. Human data collection is prone to fatigue, inconsistent judgment, and transcription errors, while traditional automated scripts lack the flexibility to deal with messy or unstructured inputs. In practice, many decisions about how data were collected or interpreted were informal and poorly documented, making it difficult for others to assess or replicate the process. Use of an LLM-driven tool emerges against this backdrop as a way to reduce these long-standing frictions, not by eliminating human judgment but by making data collection faster, more flexible, and easier to document.

10.6.3 What This Method Does

At a practical level, LLM-driven tool use helps researchers offload many of the procedural steps involved in assembling datasets. Instead of specifying every action in advance, researchers can describe their goal at a higher level and allow the model to carry out the intermediate steps, such as finding relevant sources, extracting information, and organizing outputs into structured formats. This makes it easier to move from a research idea to a usable dataset quickly.

A key strength of this approach is its ability to work across different sources and formats within a single workflow. Large language model–driven pipelines can combine web search, browsing, document parsing, and data cleaning that allow researchers to bring together information that would otherwise require multiple separate processes. This is especially useful when working with semi-structured materials like PDFs or scanned documents, which often fall between the cracks of traditional automation tools.

Another advantage is flexibility. Because the model can inspect intermediate outputs and adjust its actions, these workflows are better able to handle real-world messiness, such as inconsistent formatting or missing information. That said, the method primarily operates upstream of analysis. It changes how data are collected and prepared, not how substantive claims are ultimately made.

10.6.4 Example Applications

In practice, LLM-driven tool use is particularly useful for building structured datasets from information that is publicly available but scattered across many sources. In political science, this often includes biographical or institutional data about political elites, which may be distributed across official websites, news articles, and archival records. Tasks like compiling educational backgrounds or career histories, which traditionally required extensive manual effort, can now be completed much more quickly. As a concrete example for illustrative purposes, a modern chat-based LLM was able to collect the alma mater of all sitting US Senators and arrange this information in an extractable table format in a matter of seconds.

PROOFS

Another common application involves collecting policy documents and administrative materials. Governments and international organizations publish large volumes of reports and regulations in heterogeneous formats, often without consistent metadata. Large language model–driven workflows can automate the retrieval of these documents, extract relevant content, and standardize outputs across cases, making comparative analysis more feasible. As a further example, a modern LLM was able to extract the names of speakers and speaking times from a test set of congressional hearing transcripts and assemble the information in a standardized format, again in a matter of seconds.

These tools are also well suited for monitoring dynamic information environments, such as tracking changes to party platforms or official communications over time. Where previously recollection required restarting a time-consuming human collection cycle, a simple prompt or script can be rerun on a regular basis to ensure up-to-date and consistent information over time. Across these examples, the main benefit is not that LLMs generate new data but that they dramatically lower the cost of assembling data that already exist.

10.6.5 Validation and Inference

Using LLMs for data collection introduces new validation challenges. Because LLM-driven workflows can adapt at runtime, their behavior may vary across executions, even when instructions are held constant. Models may misidentify sources or extract incorrect information, especially in unfamiliar contexts.

To address this, researchers should pay close attention to documentation and provenance. Wherever possible, workflows should retain source links, timestamps, and raw materials, as well as logs of intermediate steps like search queries. While LLM-driven pipelines make this kind of logging feasible, it does not happen automatically and requires deliberate design.

From an inferential standpoint, it is important to separate improvements in data collection from claims about causality or measurement validity. Large language model–driven tool use makes it easier to observe and catalog information, but it does not eliminate the need to think carefully about what the resulting data actually measure. Researchers should therefore be explicit about the limits of the data and cautious in extending claims beyond what the collection process can support.

10.7 LLMs as Objects of Study (Audits and Evaluation)

The preceding sections have illuminated the ways in which political science researchers can use LLMs as tools to facilitate, improve, and expand their inquiries of substantive interest. Here, we discuss a different type of research: the study of LLMs themselves, as subjects of substantive interest.

PROOFS

Why might we do this? We describe three broad motivations: First, wanting to understand an LLM in order to understand its efficacy as a tool (the "Instrumentalist" motivation); second, wanting to understand an LLM in order to understand how it affects our world (the "Explanatory" motivation); and third, wanting to understand an LLM in order to better understand ourselves or even, perhaps, as an entity worthy of deeper understanding on its own (the "Aristotelian" motivation). These motivations are presented roughly in order of the volume of recent research fitting each paradigm.

10.7.1 Instrumentalist

By instrumentalist we mean we study LLMs to better understand how to complete some task more accurately, efficiently, or quickly. From the social scientist's perspective, this is especially relevant to tasks like the annotation of unstructured data, the generation of synthetic datasets, the lucid summary of some body of related literature, and so on. In some sense, this is the lowest-hanging fruit in the focus on LLMs as an object of study, and there is already a great deal of work on use-cases. Notice that, taken literally, one need not seek to understand the precise mechanics of an LLM to improve its performance, though it probably helps. For example, it could be that researchers discover that particular tasks are most accurately performed when hyperparameters are set to certain values or when prompts are written in specific ways, but without knowing why this matters. As an analogy, "random forests" (in the sense of Breiman 2001) were already an extraordinarily useful tool with many parameters that could be tuned and tried in specific circumstances long before deeper statistical theory arrived for these methods. In that sense, the instrumental study of LLMs is about showing the types of problems they work well for and how they might be improved, without a deeper understanding of how or why they work as they do.

Nevertheless, understanding how LLMs work can help us use them more effectively as tools for research. To continue our reference to random forests, sometime after the original Breiman work, scholars established properties of the technique in more mathematical depth (e.g., Biau 2012), which then facilitated better implementations of these tools. Such work can have instrumental underpinnings: presumably one can perfect the use of methods by establishing the science behind them. But it is also interesting per se in the sense that one wants to understand the mechanics behind how LLMs reason about problems, their failure modes, why certain parameter settings matter and others do not, and so on. There is a burgeoning related literature on how LLMs reason in ways similar or different to those of human subjects. For example, Berglund et al. (2024) note that LLMs make basic logical consistency errors that humans do not as a function of simple question ordering changes. Those authors then provide some explanation of the same. While this strand of instrumentalist research skirts the explanatory paradigm (described in Section 10.7.2), it nevertheless is still instrumentalist in its goals.

PROOFS

10.7.2 Explanatory

By this we mean studying an LLM to understand how they might shape human beliefs, behavior, and society. A major impetus for this literature is the harm that LLMs might do to our research, to users, and to society more broadly. In early work, the dominant concern was "bias" of various forms (see, e.g., Bender et al. 2021). In part, this is related to the training data for such models which Bender et al. (2021) argue is likely to overrepresent already socially dominant voices on the web and be "static" in the sense that the models are not sufficiently regularly updated such that they incorporate changes in social mores. There are other concerns, too: Those authors point out that exposure to LLM language that is itself biased can lead to psychological harm or misinformed beliefs about the world. These issues are particularly worrying to the extent that LLMs are used for "downstream" tasks – a point made by previous work on algorithmic bias in public policy (e.g., Broussard 2018; Kleinberg et al. 2018).

Today, LLMs are everywhere in online information ecosystems, summarizing search results, answering chat questions, and being actively exploited to generate mis- and disinformation (Chapter 2). They represent a dramatic labor market disruption in the form of skill-based technology change and are dramatically challenging educational conventions (Chapter 7). Against this backdrop, the motivation to understand how they work is self-evidently important, even if such an understanding need not open the black box of the model. Studies that document a left-leaning ideological bias (Motoki et al. 2024) or characterize the preference for em dashes or specific vocabulary in LLM-generated text (Kobak et al. 2025, Czuma 2026) or summarize the types of "hallucinations" these models are prone to (Huang et al. 2025) generally fall into this paradigm insofar as we care what models output in terms of how it affects other things (Salvi et al. 2025; Bai et al. 2025a; Köbis et al. 2025).

10.7.3 Aristotelian

By this we mean the study of LLMs because we think they *are* "human," or at least tell us about what it means to be human, in some meaningful way. In its purest form, this paradigm is concerned with understanding LLMs as a new type of intelligent life, worthy of its own research in the same way we treat our focus on humanity as an unquestioned object of substantive interest. This is, for now, rare in social science. But other fields have explored how "human" LLMs can plausibly be thought to be (Shanahan 2024), and the implications of this for understanding the differences between humans and LLMs (e.g., Kargupta et al. 2025). In political science, scholars have been interested in related notions of LLM preferences and beliefs. For example, Aldahoul et al. (2025) and Rozado (2024) consider the political opinions of LLMs. As with explanatory approaches discussed earlier, such studies may be instrumental in that, by understanding "how human" an LLM is, we gain some insight into

PROOFS

their optimal use for some research problems. But there is also interest per se in these objects which are neither instrumental nor explanatory: Some scholars seek to understand what the LLMs are in a deeper sense relative to humans. We would argue that these endeavors do not have obvious precedent in previous work on statistical models or techniques.

With the scope for LLMs as objects of study thus established, we turn to the more practical topics of how one might study them.

10.7.4 Auditing LLMs: Methods and Problems

The idea of studying models – here LLMs – as objects of intrinsic interest may seem unusual. But the methods for such studies are familiar, and they range from a standard battery of survey questions to randomization techniques pioneered in field experiments. In a sense, to study the output of an LLM is to "audit" it in the same way that social scientists have audited firms (Bertrand and Mullainathan 2004; Gaddis 2015), politicians (Butler and Broockman 2011; Pfaff et al. 2021), and (social) media (Ribeiro et al. 2020; Bisbee et al. 2022). At the most basic level, an audit is simply a procedure for evaluating a system's behavior. But instead of measuring the firm's hiring behavior or the politician's responding behavior or the social media platform's recommendation behavior, an audit of an LLM measures the output of an AI.

An audit experiment combines the systematic measurement of an output with experimentally manipulated traits of theoretical interest, typically through a "confederate." In the example of audits of firm hiring decisions, these "confederates" are hypothetical job applicants whose resumes are identical except for the single trait of race, connoted by name. By comparing the firm's responses across applicants with black- or white-coded names, these audits shed light on racism in hiring practices. Similarly, an audit of a recommendation algorithm might use fake accounts that are designed to represent Democrat or Republican users, allowing the researcher to compare the information recommended to different users across party lines. Or an audit of an LLM might instead treat the LLM itself as something of a confederate via system prompts that are designed to elicit different types of responses. For example, Argyle et al. (2023b) and Bisbee et al. (2024) instruct an LLM to role-play as a specific type of individual and express views from their perspective.

10.7.5 The Method

In the context of studying an LLM, an audit experiment measures some quantity of interest (typically the LLM's output, which we follow convention and denote Y) across a set of prompts designed to manipulate a trait of theoretical interest (which we denote X), holding everything else constant. In the simplest setting, there might be no trait or experimental manipulation at all, and the

PROOFS

researcher is interested only in the default behavior of the LLM: Y. For example, Rozado (2023) applied a battery of standard survey questions to probe respondent ideology to the January 2023 version of ChatGPT 3.5 turbo and concluded that the LLM is left-leaning in fourteen out of fifteen political orientation tests.

Yet even in this simplified setting, the operationalization of Y can be complicated. This is because what researchers need as input for audits is not necessarily the way that LLMs produce output. Specifically, the default behavior of most LLMs is conversational text, which can require several additional data processing steps to produce a tabular dataset amenable to downstream analyses. At that point, there are more processing questions. For example, do we annotate the textual output? Should we apply natural language processing tools, such as topic models? One solution is to constrain the output via the prompt itself, requiring that the model provide only a numeric answer or provide one corresponding to a fixed set of possible responses often asked for in a format ready-made for structured data (e.g., .json or .csv). An example of the type of prompt used to put structure on the LLM's output is provided next, although, even here, it is possible that some responses will violate the guidelines, requiring careful researcher review of the output:

```
Format the output as a csv table with the following format:
group, thermometer
```

Researcher control over the X variable is arguably less complicated, if only because the experimental design typically involves permutations of a constant prompt. This prompt might be a system prompt which governs the identity or personality the LLM adopts in responding. This is an example of a system prompt iterating over a range of characteristics of interest:

```
list(
   "role" = "system",
   "content" = stringr::str_c("You are a ", age, " year old ",
                       race, " ", gender,
                       " with a ", educ,
                       ", earning $", inc, " per year. ",
                       "You are a registered ", pid,
                       " living in the USA in 2019.")
)
```

Or it might be a "user" prompt which manipulates who the LLM believes it is interacting with, such as phrasing the same prompt from the perspective of a foreigner or a native citizen.

The ability to design prompts can create the impression of tight experimental control, but existing research suggests that LLMs might be sensitive to prompt decisions in ways that a researcher might not anticipate

PROOFS

(Berglund et al. 2024; Bisbee et al. 2024). Just as a carefully worded request to provide responses in a structured tabular format might not always work, so too might a seemingly airtight user prompt fail to evoke the desired reactions. The volatility of an LLM's reactions to prompts introduces a potentially limitless amount of researcher degrees of freedom in the construction of data, whether it be to audit data to better understand the LLM or some other type of data for which the LLM is merely a helpful tool. We return to this issue next in our replication recommendations, but it is worth underscoring that the potential for a new type of "p-hacking" (prompt-hacking) motivates our emphasis on pre-registering prompts and then documenting changes made over the course of conducting research.

With these challenges in mind, once the researcher has collected some number of $y_i \in Y$ across some permutations of $x_i \in X$, they now have a tabular dataset with which to estimate quantities of the form $E[Y|X]$.

10.7.6 Validation and Inference

A first-order question at this stage of the analysis is to evaluate the construct validity of the X and Y variables. Are we measuring what we are theoretically interested in? For example, if we want to know the ideological bias of the default ChatGPT, are we sure that the standard battery of questions used over several decades of social science research will "work"?

There are many reasons to doubt that the answer is "yes," even if we believe the instrument has construct validity when applied to humans. Like human respondents, LLMs may exhibit systematic response tendencies that are unrelated to the construct a researcher intends to measure. For example, research has documented sensitivity to the ordering of response options (Pezeshkpour and Hruschka 2024), acquiescence-like response effects (Braun 2025), and social desirability bias (Salecha et al. 2024). Thus, an apparent difference in an LLM's expressed attitudes may sometimes reflect features of the measurement instrument rather than a difference in the underlying quantity of interest.

Some response patterns, however, may be less readily understood through biases familiar from human survey research. Boelaert et al. (2025) describe one such phenomenon as "machine bias." Comparing LLM-generated responses with human responses to opinion surveys, they find that models tend to produce a relatively narrow range of answer distributions and that their errors vary idiosyncratically across questions rather than consistently favoring particular social groups. In other words, the model may systematically depart from human response patterns without those departures corresponding to a recognizable human response bias or a stable social bias.

This distinction poses a particular challenge for construct validity. We have well-developed frameworks for identifying response processes that can distort measurement among human respondents, including acquiescence, social desirability, and question-order effects. LLMs may exhibit analogous biases, but they

PROOFS

may also generate systematic response patterns arising from properties of the model that have no straightforward human analogue. Consequently, instruments validated on human respondents cannot automatically be assumed to measure the same construct when administered to an LLM.

Machine bias also threatens other types of validity. For example, it is unclear whether machine bias undermines the necessary assumptions for internally valid estimates, such as SUTVA (Stable Unit Treatment Value Assumption). Do LLMs "remember" previous treatment conditions? Do LLMs know about existing manipulations from other published work? These problems will only grow as training datasets start to incorporate the explosion of research on LLMs themselves. As a preliminary solution, the researcher should include the types of post-survey questions that probe the same issues with human samples (i.e., "What do you think this survey was about?").

The validity concerns also extend to statistical validity, which captures the inferential properties of data generated by an audit of an LLM. While some LLMs have a "temperature" parameter that governs how much variation a single prompt might generate over repeated submissions, not all do. Furthermore, even when this parameter is increased, Bisbee et al. (2024) show that the variation in LLM output is vastly smaller than that observed among real human samples. Finally, it is not clear what variation should mean when examining the output of an LLM, but it is clearly context dependent. If the output of interest is a synthetic dataset meant to replicate the properties of a survey of real humans, then the lack of variance is a problem. But if the researcher merely wants to make a statement about the biases of an LLM, then it is perhaps helpful that they can get the same results – more or less – over repeated prompts. All else equal, if the quantity of interest is the LLM itself, the second or higher order moments of a distribution of output might be of substantive import.

This brings us to the last dimension of validity: external validity (and specifically, temporal validity). As demonstrated in Barrie et al. (2024), Bisbee et al. (2024), and others, the same prompt run on the same LLM might not reproduce the same results, even with the temperature parameter set to zero. There is a subtle dilemma here. On the one hand, LLM-generated output might be too stable to capture higher order moments of interest to a scholar wishing to use it to replace or substitute for humans. But on the other hand, LLM-generated output might be too unstable to adhere to discipline norms of reproducible research. For example, the current norm in the top political science journals is to provide the data and scripts necessary to reproduce all tables and figures exactly as they appear. Should we, as a discipline, decide that LLM-generated data adhere to the same requirements? As discussed in Barrie et al. (2024), this invites a broader debate over what reproducible research should be.

More broadly, the pace of change in LLMs is astonishingly rapid. There is already a generation of publications that is likely so out-of-date as to be useless, despite being published since 2023. Should we still be skeptical of synthetic datasets, or have improvements in the LLMs meant that we can now reliably

PROOF

ask an AI to role-play as some hard-to-reach subsample to fill out a survey? Relatedly, is it still the case that ChatGPT is left-leaning, or have OpenAI's promises to remove these biases been effectively executed?[4] On the one hand, this rapid rate of change in the LLM as the quantity of interest means there is always room to reaudit the algorithms, as in Motoki et al. (2024). But on the other hand, it risks an overproduction of scientific knowledge with a shelf-life of scarcely a few years (Munger 2019).

10.7.7 Example Applications

Having described the study of LLMs as an object of interest, we conclude with a nonexhaustive summary of recent research organized by the paradigms we propose. We not only summarize a selection of studies falling into each paradigm here but emphasize that this is only a sample of a flourishing literature that covers political science, economics, psychology, sociology, and more.

(1) *Instrumentalist*: The most common type of research in this paradigm investigates the usefulness of LLMs to augment or replace costly human participants. The tasks evaluated include the annotation of data (Gilardi et al. 2023; Heseltine and Clemm von Hohenberg 2024; Choi et al. 2026; Törnberg 2025); the replacement of human survey respondents (Argyle et al. 2023b; Bisbee et al. 2024); and the performance of employees (Noy and Zhang 2023; Brynjolfsson et al. 2025), just to name a few examples. An especially sobering piece of research falling under this category demonstrates that agentic AI has the capabilities to answer online surveys at scale (Westwood 2025).

(2) *Explanatory*: A growing area of research centers on the LLM as the quantity of substantive interest, both in what models produce and in how that output shapes human beliefs and behavior. Early research demonstrated ideological biases in popular LLMs (Motoki et al. 2024; Hartmann et al. 2023) and peculiarities in how these models hallucinate along with certain styles of writing (Liang et al. 2025), and sensitivity to prompts (Berglund et al. 2024, Sclar et al 2024). While these publications largely treat the output of the LLM as the quantity of interest, a separate strand of explanatory research uses LLMs as a predictor to better understand how they affect human beliefs, behavior, and the changes to social equilibria that follow. Researchers show that LLMs can generate misinformation at scale (Goldstein et al. 2024), that they can meaningfully persuade humans on political issues (Salvi et al. 2025), and that conversational AI can shift beliefs as effectively as human persuaders (Bai et al. 2025a). In economics, researchers are interested in how individuals adopt AI and how the economy as a whole

[4] https://openai.com/index/defining-and-evaluating-political-bias-in-llms/.

PROOFS

might be affected by these decisions (e.g., Bick et al. 2026; Acemoglu 2025). In other disciplines, scholars have examined AI-generated advice and shown that it can rival expert quality even on moral questions (e.g., Dillion et al. 2025) although humans may be initially reluctant to follow it (e.g., Meincke et al. 2026).

(3) *Aristotelian*: Some of this research also inhabits an Aristotelian paradigm, specifically with regard to the (potential) emergent qualities observed in recent LLMs. A more dedicated strand of this research is found in computer science, which evaluates new models on a battery of tests once thought to be the sole domain of human intelligence. An example is given by Binz and Schulz (2023), who describe how GPT-3 reasons in comparison to humans when problem-solving. Understanding how machines do or do not reason like humans is of particular interest (Kargupta et al. 2025). Of course, that an LLM behaves in ways resembling a human does not make it one. Shanahan (2024) cautions against the unreflective use of words like "knows," "believes," and "thinks" when describing these systems: such terms carry commitments about inner life that a model trained to predict the next token does not obviously satisfy, even if the shorthand is difficult to avoid in practice and may sit more comfortably on systems built differently from today's. That this care is needed at all is itself the Aristotelian point. The question is not whether LLMs are people, but what disciplined attention to their human-like behavior tells us about agency, responsibility, and the future of social interaction.

10.8 REPLICATION AND TRANSPARENCY

Replication in LLM-based research remains an unsettled methodological question, in part because researchers are drawing on different underlying standards of what replication ought to mean. One view, associated with recent critiques of current practice, holds that replication standards for LLM-based research are often under-specified and too weak (see, e.g., Barrie et al. 2024). On this view, many current workflows rely too heavily on proprietary systems, unstable endpoints, and incomplete reporting, leaving readers with little more than an invitation to trust the researcher's description of what the model did. A stronger standard would demand something closer to full computational reproducibility wherever possible, including the use of open-weight, version-controlled, and archivable models that can be rerun under fixed conditions. In at least some cases, that aspiration is technically plausible: Unlike commercial API endpoints, locally controlled model artifacts are not vulnerable to provider-side deprecation, and some open-weight inference environments can be configured in ways that substantially improve reproducibility relative to default hosted inference.

PROOFS

A second view reaches a different conclusion. On this account, LLM-based workflows are not fundamentally different from other accepted forms of stochastic or variable data production in social science – including crowd work, survey response, human coding, or research assistance. Human coders do not produce perfectly identical labels across time; researchers do not write identical treatments when asked to generate them twice; and no one expects data collection in changing environments to yield exact replay at the record level. From this perspective, the relevant question is not whether a workflow can always reproduce the same output string, but whether it is documented clearly enough to audit and stable enough to support the same substantive inference. What matters is not perfect determinism but whether the method is transparent, inspectable, and robust under reasonable variation.

Authors on this chapter take different positions on these questions. What we agree on is that the absolute minimum standard is full workflow documentation. Whatever one's view about exact reproducibility, no LLM-based result should rest on an undocumented pipeline. Researchers should, therefore, preserve and disclose, at a minimum, the prompts and prompt templates used, the model name and version or snapshot, the date of access, decoding parameters, tool configurations, preprocessing decisions, and the code linking model outputs to final analytic variables. Wherever possible, they should also preserve the raw outputs, logs, transcripts, caches, or trace records needed to reconstruct how a result was produced. This does not by itself guarantee exact replay, but it is the essential precondition for auditability. Note that there are ongoing efforts to formalize this kind of reporting, which may become standardized in the discipline (see, e.g., Feuerriegel et al. 2025).

The specific replication burden also depends on the method. In Section 10.2 (annotation and measurement), the authors note the key requirement is that labeling workflows be auditable and that downstream inferences remain stable even if exact record-level labels vary across reruns. In Section 10.3 (treatments and stimuli), the relevant standard is not identical wording per se, but whether the same construct is being manipulated in a consistent way and whether the estimated treatment effects replicate. In Section 10.4 (silicon sampling), the central target is not the recreation of identical synthetic respondents, but the recovery of similar distributions, relationships, and substantive patterns under fully disclosed prompting, persona construction, and analysis choices. In Section 10.5 (generative ABMs), replication requires preserving the full-model specification: prompts, agent definitions, memory structures, orchestration rules, model conditions, and trace artifacts, so that others can evaluate whether the same design yields similar aggregate outcomes even if no single narrative trajectory is reproduced exactly. In Section 10.6 (tool use and data collection), replication centers on process transparency: Researchers must document the tools used, retrieval steps, decision rules, timestamps, and intermediate artifacts so that others can inspect the workflow and rerun it under

PROOFS

comparable conditions, even when the external information environment has changed. And in Section 10.7 (LLMs as objects of study), the key requirement is that prompts, benchmarks, evaluation procedures, and model conditions be specified precisely enough that others can test whether the reported claims about model behavior are robust.

The practical implication is that replication in LLM-based research should usually be evaluated at the level of substantive claims, not only at the level of exact textual replay. Stronger standards – such as frozen open-weight models and fully replayable pipelines – should be adopted whenever they are feasible and substantively important. But even where exact replay is unattainable, rigorous research is still possible if workflows are fully documented, intermediate artifacts are preserved, and claims are shown to survive defensible changes in prompts, model versions, sampling choices, or execution conditions. In that sense, the right standard for LLM-based research is neither "anything goes" nor "exact replay or failure." It is a standard of explicit documentation first, and method-appropriate robustness second.

10.9 DISCUSSION

Sections 10.1–10.8 have shown that LLMs are not merely new tools for accelerating familiar workflows. They also raise foundational questions for how political scientists define measurement, validate evidence, and understand the objects of inquiry in the first place. We conclude by highlighting five broader issues that cut across the empirical applications discussed in this chapter.

First, LLMs increasingly challenge the status of the human benchmark. Many established research workflows in political science rest on the assumption that human judgment is the appropriate standard against which machine performance should be evaluated. Yet in some domains – especially annotation and classification – LLMs may now match or exceed the consistency, speed, and sometimes even the accuracy of human coders. This raises a basic methodological question: If the model appears to outperform the human annotator, what becomes of the traditional "gold standard"? The likely implication is not that human validation disappears, but that validation must be reconceived. Rather than treating human judgment as a final ground truth, researchers may need to rely more heavily on adjudication, triangulation across coders and models, and closer attention to construct validity.

Second, the field is operating under deep uncertainty about the future trajectory of model capabilities. It remains unclear whether current systems represent an early plateau, a temporary transitional stage, or the beginning of much more substantial capability gains. This uncertainty matters methodologically because it makes today's best practices provisional. A workflow that seems fragile today may become routine tomorrow; conversely, a current use case may later appear misguided once model behavior changes or becomes better

PROOFS

Tucker_9781047774734_C010.indd 352 07-Oct-26 6:35:55 AM

understood. Political methodologists should, therefore, avoid hard-coding assumptions about what LLMs can or cannot do; instead, they should build research designs that can accommodate rapid shifts in model performance.

Third, access to LLMs is itself an issue of methodological inequality. The ability to use state-of-the-art models depends on resources: funding for API access, institutional computing infrastructure, technical expertise, and time to build and validate workflows. These resources are unevenly distributed across universities, departments, and national contexts. As a result, the spread of LLM-based methods may widen existing inequalities in who can conduct computationally intensive research, at what scale, and under what standards of reproducibility. Open-weight models and shared infrastructure may mitigate some of these disparities, but they do not eliminate them. Questions of access are, therefore, not peripheral to methodology; they shape who can participate in methodological change and on what terms.

Fourth, LLM-based research is constrained by opacity. For many of the most capable systems, researchers do not have access to the training data, model weights, post-training procedures, or provider-side interventions that shape outputs. This creates a tension at the heart of scientific inquiry. Political scientists are being asked to incorporate into their workflows systems whose internal operation is often only partially observable and whose behavior may change over time. Even when these models are useful, they remain difficult to interpret mechanistically and difficult to reproduce exactly. For that reason, the chapter has emphasized auditability, documentation, and robustness checks as minimum standards. But the broader problem remains: Methodological dependence on opaque proprietary systems places limits on cumulative scientific knowledge.

Finally, LLMs have become worthy objects of study in their own right. Political scientists are using them not only to annotate texts, generate treatments, or simulate respondents; they are increasingly studying the models themselves – their biases, values, reasoning patterns, and social effects. This development marks a significant expansion in the scope of political methodology. Large language models are now simultaneously instruments, environments, and subjects. They shape information exposure, mediate communication, and may influence beliefs and behavior at scale. For that reason, understanding how they operate, how stable their properties are, and how they interact with political life is no longer simply a technical matter. It is also a substantive problem for the discipline.

Taken together, these issues suggest that the rise of LLMs is not just another episode of methodological innovation. It is a moment that forces political science to revisit some of its most basic assumptions: what counts as valid measurement, what replication should mean, what forms of evidence are acceptable, and what kinds of entities belong within the scope of political inquiry. The challenge is not to decide once and for all whether LLMs are good or bad for the discipline. It is to develop standards that allow researchers to use them productively, critique them rigorously, and study them seriously as part of a rapidly changing methodological landscape.

PROOFS

BIBLIOGRAPHY

Acemoglu, Daron. 2025. "The Simple Macroeconomics of AI." *Economic Policy* 40(121): 13–58.

Aguinis, Herman, and Kyle J. Bradley. 2014. "Best Practice Recommendations for Designing and Implementing Experimental Vignette Methodology Studies." *Organizational Research Methods* 17(4): 351–371.

Aher, Gati V., Rosa I. Arriaga, and Adam Tauman Kalai. 2023. "Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies." In *Proceedings of the International Conference on Machine Learning*, vol. 202, 337–371. PMLR.

Ahluwalia, Aman, and Suhrud Wani. 2024. "Leveraging Large Language Models for Web Scraping." *arXiv* preprint arXiv:2406.08246.

Aldahoul, Nouar, Hazem Ibrahim, Matteo Varvello, Aaron Kaufman, Talal Rahwan, and Yasir Zaki. 2025. "Large Language Models Are Often Politically Extreme, Usually Ideologically Inconsistent, and Persuasive Even in Informational Contexts." *arXiv* preprint arXiv:2505.04171.

Anagnostidis, Sotiris, and Jannis Bulian. 2024. "How Susceptible Are LLMs to Influence in Prompts?" *arXiv* preprint arXiv:2408.11865.

Argyle, Lisa P., Christopher A. Bail, Ethan C. Busby, Joshua R. Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. 2023a. "Leveraging AI for Democratic Discourse: Chat Interventions Can Improve Online Political Conversations at Scale." *Proceedings of the National Academy of Sciences* 120(41): e2311627120.

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023b. "Out of One, Many: Using Language Models to Simulate Human Samples." *Political Analysis* 31(3): 337–351.

Argyle, Lisa P., Ethan C. Busby, Joshua R. Gubler, Bryce Hepner, Alex Lyman, and David Wingate. 2025a. "Arti-fickle Intelligence: Using LLMs as a Tool for Inference in the Political and Social Sciences." *Nature Computational Science* 5(9): 737–744.

Argyle, Lisa P., Ethan C. Busby, Joshua R. Gubler, Alex Lyman, Justin Olcott, Jackson Pond, and David Wingate. 2025b. "Testing Theories of Political Persuasion Using AI." *Proceedings of the National Academy of Sciences* 122(18): e2412815122.

Ashery, Ariel Flint, Luca Maria Aiello, and Andrea Baronchelli. 2025. "Emergent Social Conventions and Collective Bias in LLM Populations." *Science Advances* 11(20): eadu9368. https://doi.org/10.1126/sciadv.adu9368.

Ashokkumar, Ashwini, Luke Hewitt, Isaias Ghezae, and Robb Willer. 2026. "Large Language Models Can Predict the Results of Social Science Experiments." *Nature*. https://doi.org/10.1038/s41586-026-10742-x. Advance online publication.

Axelrod, Robert. 1997. *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. Princeton, NJ: Princeton University Press.

Bai, Hui, Jan G. Voelkel, Shane Muldowney, Johannes C. Eichstaedt, and Robb Willer. 2025a. "LLM-Generated Messages Can Persuade Humans on Policy Issues." *Nature Communications* 16(1): 6037.

PROOFS

Bai, Xuechunzi, Angelina Wang, Ilia Sucholutsky, and Thomas L. Griffiths. 2025b. "Explicitly Unbiased Large Language Models Still Form Biased Associations." *Proceedings of the National Academy of Sciences* 122(8): e24162281 22.

Bail, Christopher A. 2024. "Can Generative AI Improve Social Science?" *Proceedings of the National Academy of Sciences* 121(21): e2314021121. https://doi.org/10.1073/pnas.2314021121.

Barrie, Christopher, Alexis Palmer, and Arthur Spirling. 2025. *Replication for Language Models: Problems, Principles, and Best Practices for Political Science*. Manuscript, November 4. https://arthurspirling.org/documents/BarriePalmerSpirling_TrustMeBro.pdf.

Barrie, Christopher, and Petter Törnberg. 2025. "Emergent LLM Behaviors Are Observationally Equivalent to Data Leakage." *arXiv* preprint arXiv:2505.23796.

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.

Berglund, Lukas, et al. 2024. eversal Curse: LLMs trained on 'A is B' fail to learn 'B is A'." *International Conference on Learning Representations*. Vol. 2024.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review* 94(4): 991–1013.

Biau, Gérard. 2012. "Analysis of a Random Forests Model." *Journal of Machine Learning Research* 13(1): 1063–1095.

Binz, Marcel, and Eric Schulz. 2023. "Using Cognitive Psychology to Understand GPT-3." *Proceedings of the National Academy of Sciences* 120(6): e2218523120.

Bick, Alexander, Adam Blandin, and David J. Deming. 2026 "The Rapid Adoption of Generative AI." *Management Science*.

Binz, Marcel, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Eltető, et al. 2025. "A Foundation Model to Predict and Capture Human Cognition." *Nature* 644(8078): 1002–1009. https://doi.org/10.1038/s41586-025-09215-4. Advance online publication.

Bisbee, James, Megan Brown, Angela Lai, Richard Bonneau, Joshua A. Tucker, and Jonathan Nagler. 2022. "Election Fraud, YouTube, and Public Perception of the Legitimacy of President Biden." *Journal of Online Trust and Safety* 1(3).

Bisbee, James, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. "Synthetic Replacements for Human Survey Data? The Perils of Large Language Models." *Political Analysis* 32(4): 401–416.

Boelaert, Julien, Samuel Coavoux, Etienne Ollion, Ivaylo Petev, and Patrick Präg. 2025. "Machine Bias: How Do Generative Language Models Answer Opinion Polls?" *Sociological Methods & Research* 54(3): 1156–1196. https://doi.org/10.1177/00491241251330582 . Advance online publication.

Bol, Damien, and Pierre-Henri Bono. 2025. "Can ChatGPT Accurately Identify the Position of Parties? A Validation Study with an Expert Survey in France." *Research & Politics* 12 (2): 20531680251335653.

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. "Man Is to Computer Programmer as Woman Is to Homemaker?

PROOFS

Debiasing Word Embeddings." *Advances in Neural Information Processing Systems* 29: 4349–4357.
Bonabeau, Eric. 2002. "Agent-Based Modeling: Methods and Techniques for Simulating Human Systems." *Proceedings of the National Academy of Sciences* 99(suppl. 3): 7280–7287. https://doi.org/10.1073/pnas.082080899.
Bowman, Samuel R. 2023. "Eight Things to Know About Large Language Models." *arXiv* preprint arXiv:2304.00612.
Braun, D., 2025, September. "Acquiescence Bias in Large Language Models." In *EMNLP* (Findings) (pp. 11341–11355).
Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45(1): 5–32.
Briggs, Ryan, Jonathan Mellon, Vincent Arel-Bundock, and Tim Larson. 2025. "We Used LLMs to Track Methodological and Substantive Publication Patterns in Political Science and They Seem to Do a Pretty Good Job." *OSF Preprint*. https://osf.io/v7fe8.
Broska, David, Michael Howes, and Austin van Loon. 2025. "The Mixed Subjects Design: Treating Large Language Models as Potentially Informative Observations." *Sociological Methods & Research* 54(3): 1074–1109. https://doi.org/10.1177/00491241251326865. Advance online publication.
Broussard, Meredith. 2018. *Artificial Unintelligence: How Computers Misunderstand the World*. Cambridge, MA: MIT Press.
Brucks, Melanie, and Olivier Toubia. 2025. "Prompt Architecture Induces Methodological Artifacts in Large Language Models." *PLOS ONE* 20(4): e0319159.
Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond. 2025. "Generative AI at Work." *Quarterly Journal of Economics* 140(2): 889–942.
Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. "Sparks of Artificial General Intelligence: Early Experiments with GPT-4." *arXiv* preprint arXiv:2303.12712.
Butler, Daniel M., and David E. Broockman. 2011. "Do Politicians Racially Discriminate against Constituents? A Field Experiment on State Legislators." *American Journal of Political Science* 55(3): 463–477.
Butler, Daniel M., and Charles Crabtree. 2021. "Audit Studies in Political Science." In *Advances in Experimental Political Science*, edited by James N. Druckman and Donald P. Green, 42–55. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108777919.005.
Carlson, David, and Jacob M. Montgomery. 2017. "A Pairwise Comparison Framework for Fast, Flexible, and Reliable Human Coding of Political Texts." *American Political Science Review* 111(4): 835–843.
Cederman, Lars-Erik. 2003. "Modeling the Size of Wars: From Billiard Balls to Sandpiles." *American Political Science Review* 97(1): 135–150. https://doi.org/10.1017/S0003055403000571.
Cheng, Zhoujun, Jungo Kasai, and Tao Yu. 2023. "Batch Prompting: Efficient Inference with Large Language Model APIs." *arXiv* preprint arXiv:2301.08721.
Choi, Dahyun, Denis Peskoff, and Brandon Stewart. 2026. "Fine-Tuned Large Language Models Can Replicate Expert Coding Better than Trained Coders: A Study on Informative Signals Sent by Interest Groups." *Political Science Research and Methods*, advance online publication. https://doi.org/10.1017/psrm.2025.10086.
Chuang, Yun-Shiuan, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. 2024.

PROOFS

"Simulating Opinion Dynamics with Networks of LLM-Based Agents." In *Findings of the Association for Computational Linguistics: NAACL 2024*, 3326–3346. Mexico City: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-naacl.211.

Cui, Ziyan, Ning Li, and Huaikang Zhou. 2025. "A Large-Scale Replication of Scenario-Based Experiments in Psychology and Management Using Large Language Models." *Nature Computational Science* 5(8): 627–634.

Czuma, P., 2026. Em-ergence of the em-dash: a population-level rise in em-dash frequency in medRxiv preprints at the dawn of the large-language-model era. arXiv preprint arXiv:2606.29540.

Dafoe, Allan, Baobao Zhang, and Devin Caughey. 2018. "Information Equivalence in Survey Experiments." *Political Analysis* 26(4): 399–416.

Damann, Taylor J., Dean Knox, and Christopher Lucas. 2025. "A Framework for Studying Causal Effects of Speech Style: Application to US Presidential Campaigns." *Journal of the Royal Statistical Society: Series A: Statistics in Society* 189(2): 1184–1203. https://doi.org/10.1093/jrsssa/qnaf059. Advance online publication.

de Marchi, Scott. 1999. "Adaptive Models and Electoral Instability." *Journal of Theoretical Politics* 11(3): 393–419. https://doi.org/10.1177/0951692899011003007.

de Marchi, Scott, and Scott E. Page. 2014. "Agent-Based Models." *Annual Review of PoliticalScience* 17(1): 1–20.https://doi.org/10.1146/annurev-polisci-080812-191558.

Dillion, Danica, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. "Can AI Language Models Replace Human Participants?" *Trends in Cognitive Sciences* 27(7): 597–600.

Dillion, D., Mondal, D., Tandon, N., and Gray, K., 2025. "AI Language Model Rivals Expert Ethicist in Perceived Moral Expertise." *Scientific Reports*, 15(1): 4084.

Duarte, André V., Xuying Li, Bin Zeng, Arlindo L. Oliveira, Lei Li, and Zhuo Li. 2025. "Recap: Reproducing Copyrighted Data from LLMs' Training with an Agentic Pipeline." *arXiv* preprint arXiv:2510.25941.

Egami, Naoki, Musashi Hinck, Brandon Stewart, and Hanying Wei. 2023. "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.

Egami, Naoki, Musashi Hinck, Brandon M. Stewart, and Hanying Wei. 2024. "Using Large Language Model Annotations for the Social Sciences: A General Framework of Using Predicted Variables in Downstream Analyses." Working paper, November 17. https://naokiegami.com/paper/dsl_ss.pdf.

Epstein, Joshua M., and Robert Axtell. 1996. *Growing Artificial Societies: Social Science from the Bottom Up*. Cambridge, MA: MIT Press.

Evsyukova, Yulia, Felix Rusche, and Wladislaw Mill. 2025. "LinkedOut? A Field Experiment on Discrimination in Job Network Formation." *Quarterly Journal of Economics* 140(1): 283–334.

Feuerriegel, Stefan, Christopher Barrie, Laura Globig, Dan-Mircea Mirea, Manoel Horta Ribeiro, Steve Rathje, Arthur Spirling, and Diyi Yang. 2025. "GUIDE-LLM: A Consensus-Based Reporting Checklist for Large Language Models in Behavioral and Social Science." Working paper. https://sfeuerriegel.github.io/llm-checklist/.

Gaddis, S. Michael. 2015. "Discrimination in the Credential Society: An Audit Study of Race and College Selectivity in the Labor Market." *Social Forces* 93(4): 1451–1479.

PROOF

Gallegos, Isabel O., Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. "Bias and Fairness in Large Language Models: A Survey." *Computational Linguistics* 50(3): 1097–1179.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. 2023. "ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks." *Proceedings of the National Academy of Sciences* 120(30): e2305016120.

Goldstein, Josh A., Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. 2024. "How Persuasive Is AI-Generated Propaganda?" *PNAS Nexus* 3(2): pgae034.

Goli, Ali, and Amandeep Singh. 2024. "Can Large Language Models Capture Human Preferences?" *Marketing Science* 43(4): 709–722.

Götz, Friedrich M., Rakoen Maertens, Sahil Loomba, and Sander van der Linden. 2024. "Let the Algorithm Speak: How to Use Neural Networks for Automatic Item Generation in Psychological Scale Development." *Psychological Methods* 29(3): 494.

Grimm, Volker, Steven F. Railsback, Christian E. Vincenot, Uta Berger, Cara Alyse Gallagher, Donald L. DeAngelis, Bruce Edmonds, Jiaqi Ge, Jarl Giske, Jürgen Groeneveld, Alice S. A. Johnston, Alexander Milles, Jacob Nabe-Nielsen, J. Gareth Polhill, Viktoriia Radchuk, Marie-Sophie Rohwäder, Richard A. Stillman, Jan C. Thiele, and Daniel Ayllón. 2020. "The ODD Protocol for Describing Agent-Based and Other Simulation Models: A Second Update to Improve Clarity, Replication, and Structural Realism." *Journal of Artificial Societies and Social Simulation* 23(2): 7. https://doi.org/10.18564/jasss.4259.

Grimmer, Justin. 2015. "We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together." *PS: Political Science & Politics* 48(1): 80–83.

Gui, George, and Olivier Toubia. 2023. "The Challenge of Using LLMs to Simulate Human Behavior: A Causal Inference Perspective." *arXiv* preprint arXiv:2312.15524.

Hackenburg, Kobi, and Helen Margetts. 2024. "Evaluating the Persuasive Influence of Political Microtargeting with Large Language Models." *Proceedings of the National Academy of Sciences* 121(24): e2403116121.

Hartmann, Jochen, Jasper Schwenzow, and Maximilian Witte. 2023. "The Political Ideology of Conversational AI: Converging Evidence on ChatGPT's Pro-Environmental, Left-Libertarian Orientation." *arXiv* preprint arXiv:2301.01768.

Heseltine, Michael, and Bernhard Clemm von Hohenberg. 2024. "Large Language Models as a Substitute for Human Experts in Annotating Political Text." *Research & Politics* 11(1): 20531680241236239.

Horton, John J. 2023. "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?" NBER Working Paper No. 31122. Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w31122.

Huang, Lei, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions." *ACM Transactions on Information Systems* 43(2): 1–55.

Huang, Yupan, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. "LayoutLMv3: Pre-Training for Document AI with Unified Text and Image Masking." *arXiv* preprint arXiv:2204.08387. https://doi.org/10.1145/3503161.3548112.

PROOFS

Hullman, Jessica, David Broska, Huaman Sun, and Aaron Shaw. 2025. "This Human Study Did Not Involve Human Subjects: Validating LLM Simulations as Behavioral Evidence." *arXiv* preprint arXiv:2602.15785. https://doi.org/10.48550/arXiv.2602.15785.

Imai, Kosuke, and Kentaro Nakamura. 2025. "GenAI-Powered Inference." *arXiv* preprint arXiv:2507.03897.

Kargupta, P., Li, S.S., Wang, H., Lee, J., Chen, S., Ahia, O., Light, D., Griffiths, T.L., Kleiman-Weiner, M., Han, J. and Celikyilmaz, A., 2025. "Cognitive Foundations for Reasoning and Their Manifestation in LLMs." *arXiv* preprint arXiv:2511.16660.

Kim, Geewook, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2021. "OCR-Free Document Understanding Transformer." *arXiv* preprint arXiv:2111.15664.

Kim, Junsol, and Byungkyu Lee. 2023. "AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction." *arXiv* preprint arXiv:2305.09620.

Kirk, Robert, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2023. "Understanding the Effects of RLHF on LLM Generalisation and Diversity." *arXiv* preprint arXiv:2310.06452.

Kittur, Aniket, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. "The Future of Crowd Work." In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 1301–1318. San Antonio, TX, February 23–27. New York: Association for Computing Machinery. https://doi.org/10.1145/2441776.2441923.

Kiyomaru, Hirokazu, Issa Sugiura, Daisuke Kawahara, and Sadao Kurohashi. 2024. "A Comprehensive Analysis of Memorization in Large Language Models." In *Proceedings of the 17th International Natural Language Generation Conference*, 584–596. Tokyo, Japan: Association for Computational Linguistics. https://aclanthology.org/2024.inlg-main.45/

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

Kobak, Dmitry, Rita González-Márquez, Emőke-Ágnes Horvát, and Jan Lause. 2025. "Delving into LLM-Assisted Writing in Biomedical Publications through Excess Vocabulary." *Science Advances* 11(27): eadt3813.

Köbis, Nils, Zoe Rahwan, Raluca Rilla, Bramantyo Ibrahim Supriyatno, Clara Bersch, Tamer Ajaj, Jean-François Bonnefon, and Iyad Rahwan. 2025. "Delegation to Artificial Intelligence Can Increase Dishonest Behaviour." *Nature* 646(8083): 126–134. https://doi.org/10.1038/s41586-025-09505-x. Advance online publication.

Kojima, Takeshi, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. "Large Language Models Are Zero-Shot Reasoners." *arXiv* preprint arXiv:2205.11916.

Kozlowski, Austin C., and James Evans. 2025. "Simulating Subjects: The Promise and Peril of Artificial Intelligence Stand-Ins for Social Agents and Interactions." *Sociological Methods & Research* 54(3): 1017–1073. https://doi.org/10.1177/00491241251337316. Advance online publication.

PROOFS

Lai, Ruilin, Christopher Lucas, Jacob M. Montgomery, and Margit Tavits. 2026. "The Emotional Architecture of Populism: The Divergent Effects of Anti-Elite and People-Centric Frames." Unpublished working paper.

Landgrave, Michelangelo, and Nicholas Weller. 2022. "Do Name-Based Treatments Violate Information Equivalence? Evidence from a Correspondence Audit Experiment." *Political Analysis* 30(1): 142–148.

Larooij, Maik, and Petter Törnberg. 2026. "Validation Is the Central Challenge for Generative Social Simulation: A Critical Review of LLMs in Agent-Based Modeling." *Artificial Intelligence Review* 59(1): 15. https://doi.org/10.1007/s10462-025-11412-6.

Lasswell, Harold D. 1941. "The World Attention Survey." *Public Opinion Quarterly* 5(3): 456–462. https://doi.org/10.1086/265515.

Lazer, David, and Katya Ognyanova. 2024. "The Future of Computational Social Science." In Janet M. Box-Steffensmeier, Dino P. Christenson, and Valeria Sinclair-Chapman, eds., *The Oxford Handbook of Engaged Methodological Pluralism in Political Science*, 1–22. Oxford: Oxford University Press.

Lazer, David M. J., Alex Pentland, Duncan J. Watts, Sinan Aral, Susan Athey, Noshir Contractor, Deen Freelon, Sandra Gonzalez-Bailon, Gary King, Helen Margetts, Alondra Nelson, Matthew J. Salganik, Markus Strohmaier, Alessandro Vespignani, and Claudia Wagner. 2020. "Computational Social Science: Obstacles and Opportunities." *Science* 369(6507): 1060–1062.

Liang, Weixin, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, et al. 2025. "Quantifying Large Language Model Usage in Scientific Papers." *Nature Human Behaviour* 9(12): 2599–2609. https://doi.org/10.1038/s41562-025-02273-8 Advance online publication.

Lu, Yikang, Alberto Aleta, Chunpeng Du, Lei Shi, and Yamir Moreno. 2024. "LLMs and Generative Agent-Based Models for Complex Systems Research." *Physics of Life Reviews* 51: 283–293. https://doi.org/10.1016/j.plrev.2024.10.013.

Luo, Yun, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2025. "An Empirical Study of Catastrophic Forgetting in Large Language Models during Continual Fine-Tuning." *IEEE/ACM Transactions on Audio, Speech, and Language Processing, advance*, online publication.

Lyman, Alex, Bryce Hepner, Lisa P. Argyle, Ethan C. Busby, Joshua R. Gubler, and David Wingate. 2025. "Balancing Large Language Model Alignment and Algorithmic Fidelity in Social Science Research." *Sociological Methods & Research* 54(3): 1110–1155.

Macy, Michael W., and Robert Willer. 2002. "From Factors to Actors: Computational Sociology and Agent-Based Modeling." *Annual Review of Sociology* 28: 143–166.

Mao, Kelong, Zhicheng Dou, Fengran Mo, Jiewen Hou, Haonan Chen, and Hongjin Qian. 2023. "Large Language Models Know Your Contextual Search Intent: A Prompting Framework for Conversational Search." *arXiv* preprint arXiv:2303.06573.

Meincke, L., Nave, G. and Terwiesch, C., 2026. "Advice Quality and Source Disclosure Shape Trust in AI-Generated Ethical Advice." *Scientific Reports*, 16(1): 11868.

Mellon, Jonathan, Jack Bailey, Ralph Scott, James Breckwoldt, Marta Miori, and Phillip Schmedeman. 2024. "Do AIs Know What the Most Important Issue Is? Using Language Models to Code Open-Text Social Survey Responses at Scale." *Research & Politics* 11(1): 20531680241231468.

PROOFS

Motoki, Fabio, Valdemar Pinho Neto, and Victor Rodrigues. 2024. "More Human than Human: Measuring ChatGPT Political Bias." *Public Choice* 198(1): 3–23.

Mou, Xinyi, Zhongyu Wei, and Xuanjing Huang. 2024. "Unveiling the Truth and Facilitating Change: Towards Agent-Based Large-Scale Social Movement Simulation." In *Findings of the Association for Computational Linguistics: ACL 2024*, 4789–4809. Bangkok, Thailand: Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.285.

Munger, Kevin. 2019. "The Limited Value of Non-Replicable Field Experiments in Contexts with Low Temporal Validity." *Social Media + Society* 5(3): 2056305119859294.

Münker, Simon, Nils Schwager, and Achim Rettinger. 2025. "Don't Trust Generative Agents to Mimic Communication on Social Networks Unless You Benchmarked Their Empirical Realism." *arXiv* preprint arXiv:2506.21974.

Nakano, Reiichiro, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. "WebGPT: Browser-Assisted Question-Answering with Human Feedback." *arXiv* preprint arXiv:2112.09332.

Neuendorf, Kimberly A. 2016. *The Content Analysis Guidebook*. Thousand Oaks, CA: SAGE.

Noy, Shakked, and Whitney Zhang. 2023. "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence." *Science* 381(6654): 187–192.

OpenAI. 2023. "GPT-4 Technical Report." *arXiv* preprint arXiv:2303.08774.

Ornstein, Joseph T., Elise N. Blasingame, and Jake S. Truscott. 2025. "How to Train Your Stochastic Parrot: Large Language Models for Political Texts." *Political Science Research and Methods* 13(2): 264–281.

Pagan, Nicolò, Petter Törnberg, Christopher A. Bail, Anikó Hannák, and Christopher Barrie. 2025. "Computational Turing Test Reveals Systematic Differences between Human and AI Language." *arXiv* preprint arXiv:2511.04195.

Palmer, Alexis, and Arthur Spirling. 2023. "Large Language Models Can Argue in Convincing Ways about Politics, but Humans Dislike AI Authors: Implications for Governance." *Political Science* 75(3): 281–291.

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.

Park, Joon Sung, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. "Generative Agents: Interactive Simulacra of Human Behavior." In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, 1–22. https://doi.org/10.1145/3586183.3606763.

Park, Joon Sung, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. "LLM Agents Grounded in Self-Reports Enable General-Purpose Simulation of Individuals." *arXiv* preprint arXiv:2411.10109. https://doi.org/10.48550/arXiv.2411.10109.

Peng, Tianyi, George Gui, Melanie Brucks, Daniel J. Merlau, Grace Jiarui Fan, Malek Ben Sliman, Eric J. Johnson, Abdullah Althenayyan, Silvia Bellezza, Dante Donati,

PROOFS

et al. 2025. "Digital Twins as Funhouse Mirrors: Five Key Distortions." *arXiv* preprint arXiv:2509.19088. https://doi.org/10.48550/arXiv.2509.1908.

Pezeshkpour, P., and Hruschka, E. 2024, June. "Large Language Models Sensitivity to the Order of Options in Multiple-choice Questions." In *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 2006–2017).

Pfaff, Steven, Charles Crabtree, Holger L. Kern, and John B. Holbein. 2021. "Do Street-Level Bureaucrats Discriminate Based on Religion? A Large-Scale Correspondence Experiment among American Public School Principals." *Public Administration Review* 81(2): 244–259.

Piao, Jinghua, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. 2025. "AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents Advances Understanding of Human Behaviors and Society." *arXiv* preprint arXiv:2502.08691. https://doi.org/10.48550/arXiv.2502.08691.

Porter, Ethan, and Yamil R. Velez. 2022. "Placebo Selection in Survey Experiments: An Agnostic Approach." *Political Analysis* 30(4): 481–494.

Porter, Nathaniel D., Ashton M. Verdery, and S. Michael Gaddis. 2020. "Enhancing Big Data in the Social Sciences with Crowdsourcing: Data Augmentation Practices, Techniques, and Opportunities." *PLOS ONE* 15(6): e0233154.

Qu, Yao, and Jue Wang. 2024. "Performance and Biases of Large Language Models in Public Opinion Simulation." *Humanities and Social Sciences Communications* 11(1): 1–13.

Ribeiro, Manoel Horta, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, and Wagner Meira Jr. 2020. "Auditing Radicalization Pathways on YouTube." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 131–141. Barcelona, Spain, January 27–30. New York: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372879.

Rozado, David. 2023. "The Political Biases of ChatGPT." *Social Sciences* 12(3): 148.

Rozado, David. 2024. "The Political Preferences of LLMs." *PLOS ONE* 19(7): e0306621.

Salvi, Francesco, et al. 2025. "On the Conversational Persuasiveness of GPT-4." *Nature Human Behaviour* 9.8: 1645–1653.

Salecha, A., Ireland, M.E., Subrahmanya, S., Sedoc, J., Ungar, L.H. and Eichstaedt, J.C., 2024. "Large Language Models Show Human-Like Social Desirability Biases in Survey Responses." *arXiv* preprint arXiv:2405.06058.

Santurkar, Shibani, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. "Whose Opinions Do Language Models Reflect?" In *Proceedings of the 40th International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, vol. 202: 29971–30004. PMLR.

Schelling, Thomas C. 1971. "Dynamic Models of Segregation." *Journal of Mathematical Sociology* 1(2): 143–186. https://doi.org/10.1080/0022250X.1971.9989794.

Sclar, M., Choi, Y., Tsvetkov, Y. and Suhr, A., 2024, May. "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I Learned to Start Worrying about Prompt Formatting." In *International Conference on Learning Representations* (Vol. 2024, pp. 25055–25083).

Shanahan, Murray. 2024. "Talking about Large Language Models." *Communications of the ACM* 67(2): 68–79.

PROOFS

Suh, Joseph, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. 2025. "Language Model Fine-Tuning on Scaled Survey Data for Predicting Distributions of Public Opinions." *arXiv* preprint arXiv:2502.16761.

Tang, Jiakai, Heyang Gao, Xuchen Pan, Lei Wang, Haoran Tan, Dawei Gao, Yushuo Chen, Xu Chen, Yankai Lin, Yaliang Li, Bolin Ding, Jingren Zhou, Jun Wang, and Ji-Rong Wen. 2025. "GenSim: A General Social Simulation Platform with Large Language Model Based Agents." In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, 143–150. Albuquerque: Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-demo.15.

Törnberg, Petter. 2025. "Large Language Models Outperform Expert Coders and Supervised Classifiers at Annotating Political Social Media Messages." *Social Science Computer Review* 43(6): 1181–1195. https://doi.org/10.1177/08944393241286471.

Törnberg, Petter, Diliara Valeeva, Justus Uitermark, and Christopher Bail. 2023. "Simulating Social Media Using Large Language Models to Evaluate Alternative News Feed Algorithms." *arXiv* preprint arXiv:2310.05984.

Toubia, Olivier, George Z. Gui, Tianyi Peng, Daniel J. Merlau, Ang Li, and Haozhe Chen. 2025. "Database Report: Twin-2K-500: A Data Set for Building Digital Twins of over 2,000 People Based on Their Answers to over 500 Questions." *Marketing Science* 44(6): 1446–1455.

Vafa, Keyon, Sarah Bentley, Jon Kleinberg, and Sendhil Mullainathan. 2025. "What's Producible May Not Be Reachable: Measuring the Steerability of Generative Models." *arXiv* preprint arXiv:2503.17482.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." *Advances in Neural Information Processing Systems* 30: 6000–6010.

Velez, Yamil Ricardo, and Patrick Liu. 2025. "Confronting Core Issues: A Critical Assessment of Attitude Polarization Using Tailored Experiments." *American Political Science Review* 119(2): 1036–1053.

Veselovsky, Veniamin, Manoel Horta Ribeiro, Philip J. Cozzolino, Andrew Gordon, David Rothschild, and Robert West. 2025. "Prevalence and Prevention of Large Language Model Use in Crowd Work." *Communications of the ACM* 68(3): 42–47.

Waight, Hannah, Solomon Messing, Anton Shirikov, Margaret E. Roberts, Jonathan Nagler, Jason Greenfield, Megan A. Brown, Kevin Aslett, and Joshua A. Tucker. 2025. "Quantifying Narrative Similarity across Languages." *Sociological Methods & Research* 54(3): 933–983. https://doi.org/10.1177/00491241251340080. Advance online publication.

Wang, Angelina, Jamie Morgenstern, and John P. Dickerson. 2025a. "Large Language Models That Replace Human Participants Can Harmfully Misportray and Flatten Identity Groups." *Nature Machine Intelligence* 7(3): 400–411. https://doi.org/10.1038/s42256-025-00986-z. Advance online publication.

Wang, Lei, Heyang Gao, Xiaohe Bo, Xu Chen, and Ji-Rong Wen. 2025b. "Yulan-OneSim: Towards the Next Generation of Social Simulator with Large Language Models." *arXiv* preprint arXiv:2505.07581. https://doi.org/10.48550/arXiv.2505.07581.

Wang, Yu. 2024. "On Finetuning Large Language Models." *Political Analysis* 32(3): 379–383.

PROOFS

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. "Emergent Abilities of Large Language Models." *arXiv* preprint arXiv:2206.07682.

Weisberg, Herbert F. 2009. *The Total Survey Error Approach: A Guide to the New Science of Survey Research*. Chicago: University of Chicago Press.

Westwood, Sean J. 2025. "The Potential Existential Threat of Large Language Models to Online Survey Research." *Proceedings of the National Academy of Sciences* 122(47): e2518075122.

Windrum, Paul, Giorgio Fagiolo, and Alessio Moneta. 2007. "Empirical Validation of Agent-Based Models: Alternatives and Prospects." *Journal of Artificial Societies and Social Simulation* 10(2): 8.

Wu, Patrick Y., Jonathan Nagler, Joshua A. Tucker, and Solomon Messing. 2023a. "Large Language Models Can be Used to Estimate the Latent Positions of Politicians." *arXiv* preprint arXiv:2303.12057. https://doi.org/10.48550/arXiv.2303.12057.

Wu, Qingyun, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2023b. "AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation." *arXiv* preprint arXiv:2308.08155.

Xu, Yiheng, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. "LayoutLM: Pre-Training of Text and Layout for Document Image Understanding." In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1192–1200. New York: ACM. https://doi.org/10.1145/3394486.3403172.

Yin, Yidan, Nan Jia, and Cheryl J. Wakslak. 2024. "AI Can Help People Feel Heard, but an AI Label Diminishes This Impact." *Proceedings of the National Academy of Sciences* 121(14): e2319112121.

Zhou, Lingfeng, Yi Xu, Zhenyu Wang, and Dequan Wang. 2025. "FlockVote: LLM-Empowered Agent-Based Modeling for Simulating U.S. Presidential Elections." *arXiv* preprint arXiv:2512.05982. https://doi.org/10.48550/arXiv.2512.05982.

Zhou, Xuhui, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. "SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents." In *Proceedings of the International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=mM7VurbA4r.

PROOFS

11

Teaching and Learning

Political Science in the Era of AI

John Ishiyama, Christine Cahill, Jennifer De Maio, Stefan E. Kehlenbach, Sing Hui Lee, Steven Michels, Charles C. Turner, and Nicole Wu

PROOFS

Artificial intelligence (AI) is a new frontier in political science, and it is one that affects almost everything we do and, in particular, how we teach. The use of AI presents significant challenges to instructors who are interested in cultivating critical thinking, problem-solving, and communication skills for students. However, the advent of AI also opens up possibilities for improving the teaching of political science. In this chapter, we examine the attitudes of political science instructors vis-à-vis AI and teaching. We then review the growing literature that addresses AI and the teaching of political science. This includes a growing skepticism over the use of AI in the classroom but also a realization that the technology offers new possibilities for political science instructors. The purpose of this chapter is not to take sides in the debate over whether AI use is a good or bad thing for political science teaching. The goal is instead to review the variety of perspectives on the use of AI and political science and to illustrate how our colleagues in the discipline have adjusted their teaching in response to the challenges presented by AI. We hope that if an instructor does choose to use AI, this chapter will help them to devise strategies to promote the responsible and ethical use of AI by students.

Many scholars of teaching in the discipline have argued that the appearance of AI has "rendered much of our pedagogical repertoire obsolete and that we need to adjust" (Raymond, 2024). This is likely an overstatement, but it does highlight the challenges that we, as political science instructors, face. In many ways, political science instructors are left to their own devices to navigate this new teaching environment. Indeed, there is no current uniform guidance on the

The authors of this chapter acknowledge the use of ChatGPT and Claude to assist in formatting Box 11.1 and some bibliographic citations. Otherwise, AI was not used in any other way and no AI-generated text was used to draft the content of this chapter.

Tucker_9781047774734_C011.indd 365 08-Sep-26 2:36:01 PM

use of AI in teaching. As Wu and Wu (2024) point out, there are widely divergent approaches to AI in higher education, ranging from rejection to embrace of its use in the classroom. For example, Harvard Law School's "Statement on Use of AI Large Language Models" states that the use of AI "in preparing to write, or writing, academic work for courses … is prohibited unless expressly identified in writing by the instructor as an appropriate resource … ." On the other extreme, the Sandra Day O'Connor College of Law at Arizona State University explicitly permits their students to use generative AI, citing its prevalence in the legal field and the school's mission "to educate and prepare the next generation of lawyers and leaders." Others leave discretion about AI usage in the classroom to individual instructors. As Wu and Wu (2024) note, "In many ways, instructors across institutional types (e.g., R1, R2, and liberal arts colleges) are left on their own to design their own assessments and evaluate students."

To begin, it is important to discuss briefly what we mean by AI, particularly as it relates to teaching and learning. Indeed, not all AI are the same. Second, we review the principal challenges that have emerged as the result of the advent of AI. In particular, we review how these challenges relate to the recommendations regarding the development of student skills made by the 2024 American Political Science Association (APSA) Presidential Task Force Report *Rethinking Political Science Education* (APSA, 2024). Third, we review how our colleagues in the discipline have reacted to AI use in the classroom (via results from a recent APSA survey and a review of the growing number of papers and publications that have addressed AI use in the classroom). Fourth, we touch upon the ethics of AI use in the classroom (or making readers aware of the costs of using AI in the classroom before deciding to use AI). Fifth, we review cases where AI has been used in assignment design, assisting with grading, promoting digital literacy, and career preparation, providing readers with potential guidelines if they choose to engage with AI in their classrooms. We then conclude with a brief discussion of what APSA might do in terms of providing guidelines to promote the ethical and responsible use of AI in the political science classroom.

11.1 DEFINITION AND CLARIFICATIONS

What do we mean by "AI," especially as it relates to teaching and learning in political science? The first thing we need to realize is that AI is, as Emily Bender and Alex Hanna put it, "a marketing term. It doesn't refer to a coherent set of technologies" (Bender and Hanna, 2025, p. 199). Instead, the term AI today is used to refer to a broad set of technological tools and applications which can seem, or are made to seem, like they are in some way imitating human thought, action, or attitudes. Bender and Hanna have developed a typology of common types of AI. These include "Decision-Making" AI, which use statistical processes to make predictive decisions about the future;

PROOFS

"Classification" AI, which is often used to categorize pictures, develop facial recognition programs, or manage large databases of consumer information for advertisements; "Recommendation" AI, which are systems that leverage the large datasets compiled by classification AI in order to organize social media websites, recommend songs in Spotify, or predict what someone will watch next in Netflix; "Transcription/Translation" AI, which includes both automatic speech recognition as well as language translation; and, finally, "Text and Image Generation" AI, which is perhaps the most front-of-mind type of AI and includes ChatGPT and other systems that create human-seeming text or images.

In general, we can classify all these diverse types of AI into two broad categories: "Predictive AI" and "Generative AI." The category of predictive AI includes recommendation and classification algorithms that are aimed at drawing probabilistic conclusions from existing data sets – for example, facial recognition or engaging with recidivism predictions. These systems tend to be more hidden from public view because they either run in the background of entertainment apps (like Spotify's recommendation algorithm) or because they are specialty apps designed for systems like law enforcement (Brubaker, 2022). However, within education, these types of algorithms have been used for a long time as plagiarism checkers (Morris and Stommel, 2017).

When we are talking about AI impacting the teaching of political science, our language and assumptions mostly imply the use of generative AI, but there is a role for predictive AI, particularly when considering AI as a detection or grading/assessment tool.

11.2 CHALLENGES AND OPPORTUNITIES

To assess the challenges presented by AI to political science teaching and learning, it is perhaps useful to reiterate the arguments made by the APSA Presidential Task Force in its report *Rethinking Political Science Education* (APSA, 2024). It is notable that this report was written before AI exploded on the scene in 2024, and the report had little to say about how AI would affect political science education. To help frame this chapter, it merits reiterating the recommendations of that report and how AI use in the classroom affects the implementation of these recommendations.

The 2024 report argued that, facing numerous challenges to our discipline, political science needed to rethink how we conduct teaching and learning. Indeed, beyond our traditional emphasis on developing students' critical thinking, writing, and speaking skills, we must now also prepare our majors to work with, learn from, and ethically interpret data as it applies to political problems, and we must reinvigorate our teaching of civic values and civic engagement. While other social sciences share our mission of providing a traditional liberal arts education, political science uniquely instructs students about the role of political institutions, processes, and actors in the distribution of public goods and the

PROOFS

ways in which power operates through these institutions and processes in public as well as private spheres. In essence, political science aims to provide students with knowledge, skills, and values necessary for ethical and engaged citizenship at local, national, and global levels. The report highlighted not only content learning objectives but also the development of skills, including developing the ability to

(1) analyze political phenomena and critique arguments in a theoretically and empirically informed manner;
(2) engage as citizens through problem-solving, collective action, and collaborative decision-making at all levels from local to global;
(3) identify, evaluate, properly use, and cite verifiable sources of evidence;
(4) engage thoughtfully and respectfully with competing values and perspectives;
(5) use qualitative and quantitative methods effectively and appropriately; and
(6) Communicate effectively in written and verbal modes to a variety of audiences.

There are three commonly mentioned challenges presented by AI that may challenge the realization of the goals mentioned in the 2024 Task Force report: (1) Students may be able to complete and successfully pass courses without actually learning the material; (2) overreliance on AI prevents the development of critical-thinking skills and short circuits the learning process, so that students learn neither content nor skills; and (3) the question of how work is assessed. How do we assess student work when it is not their own, and how do we incentivize students to do their own work and develop important content knowledge and analytical and communication skills when they overly rely on AI? These are some of the central challenges our discipline will need to address.

11.3 SURVEYING ATTITUDES

How have political science instructors reacted to the advent of AI? And what do they see as the challenges AI presents? Wu and Wu (2024) conducted two surveys in collaboration with the APSA between April and August 2023 to examine how political science educators perceive and respond to large language models (LLMs). The first survey was embedded within APSA's annual membership survey, distributed to approximately 11,000 current and recent members. This membership survey contained four focused questions and yielded 1,615 responses. The respondent pool included faculty members (62.6 percent), graduate students (11.8 percent), adjunct or visiting professors (4.6 percent), postdoctoral fellows (3.7 percent), and the remainder in the "other" category; and 81.3 percent of the total based primarily in the United States. The second survey, the Teaching and Learning Conference (TLC) survey, targeted a subset of 3,442 individuals who had participated in APSA's TLC within the previous five years. The survey received 198 responses. This sample

PROOFS

represents educators particularly engaged with pedagogical innovation. The TLC survey included more detailed questions (both closed- and open-ended questions) about planned instructional changes, as well as a component that requested respondents to identify bot- and human-produced essays from an essay bank.

These surveys revealed several patterns in how political science educators view and respond to generative AI tools: first, despite concerns about AI, political science educators maintain that written assignments – the form of assignment with its integrity most directly threatened by LLMs (for now) – serve critical pedagogical functions. An overwhelming majority of TLC respondents indicated that written work is important for upper-level and lower-level students alike. Respondents identified improving writing skills, developing argumentative capacity, and generating original arguments as the most important goals of written assignments. At the time of the survey, written assignments comprised 62.1 percent of grades in upper-level and 45.5 percent in lower-level courses among TLC respondents.

Second, educators expressed generally negative attitudes toward AI's impact on political science education. Of colleagues surveyed, 44.9 percent rated AI tools as "bad for [their] political science classes" and only 13.6 percent rated them positively. Qualitative responses revealed concerns about independent thinking and the propagation of biases and misinformation through LLM outputs. Despite this overall skepticism, educators distinguished sharply between different uses of AI tools. They opposed using AI to write complete essays, answer exam questions, or generate ideas and outlines, but they supported using AI in an assistive manner to copy edit and format essays.

Third, behaviorally, in 2023, most planned responses focused on enforcement rather than integration. More than 30 percent of TLC respondents indicated they would decrease the number of written assignments, while 45 percent planned no change. The majority indicated that they would run assignments through LLMs themselves and implement AI-detection systems. Wu and Wu (2024) asked respondents to identify whether essays were student- or bot-written; respondents achieved accuracy rates between 45 percent and 53 percent: effectively, no better than chance. This observation, combined with high false-positive rates of detection systems, should give us pause before leaning too heavily on policing and prevention as a key strategy. Many universities disabled AI detection features citing concerns about accuracy and bias.[1]

Since the "breakout year" of 2023 for public adoption of generative AI, institutions and educators have continued to grapple with how to respond to generative AI in educational settings. Most political science departments

[1] Ghaffary (2023).

PROOFS

have not issued formal department-level policies, instead leaving AI decisions to individual instructors. A 2024 survey by *The Daily Princetonian* found that, among all Princeton departments offering undergraduate majors, only Economics had developed department-specific AI guidelines.[2] More systematically, reviewing AI policies of twenty-seven top-ranked universities in the United States (as ranked by the *U.S. News and World Report* and *Times Higher Education*), Alqahtani and Wafula (2025), focusing on materials published by university-wide teaching and learning hubs, found that faculty in a majority of universities could freely decide how AI tools are employed in their classrooms. Interestingly, they noted that most institutions "encouraged its use" (p. 673). This positive stance toward AI aligns with the aforementioned discipline-wide survey finding that 73 percent of political science educators stated that it was at least moderately important that students learn how to use generative AI tools (Wu and Wu, 2024). Indeed, discussions around this topic should not proceed as if students are encountering AI tools for the first time in college. College students today would have already navigated varying AI policies in high school, developed their own practices around AI use, and formed habits – both productive and problematic – before arriving on campus. The question confronting higher education is not whether to introduce students to AI but rather how to work with students who are already AI users, helping them develop more sophisticated approaches to these tools. In the following sections, we suggest some ways in which AI tools can be used to enhance student learning.

11.4 A LOOK AT THE LITERATURE

In recent years, there have been dozens of journal articles, popular press pieces, working papers, blogs, and conference papers addressing the use of AI in teaching political science through the end of 2025. Most of these pieces focus on generative AI (hereafter AI for the sake of simplicity). Methods in these pieces vary from thought pieces to best practices, from broad surveys to systematic empirical studies. In this section we briefly examine forty-two of these papers and classify them in terms of the following thematic categories: detection, preventing or discouraging use, ethical and equitable use/nonuse, career preparation, faculty development and incorporation, and department/class policy.

We note that the significance of AI has grown so rapidly that the most recent previous APSA Presidential Task Force Report mentions it only fleetingly despite its focus on the future of political science education (APSA, 2024). One way to measure the growth in political scientists' attention to the role of AI in teaching and learning issues is via a simple count over time. Table 11.1 indicates the number of articles/papers addressing AI and teaching over the past

[2] Sanchez and Koukas (2024).

PROOFS

TABLE 11.1 *Number of articles and papers addressing generative AI per year, in primary APSA outlets on teaching and learning*

Outlet	2022	2023	2024	2025
APSA Annual Meeting	0	5	2	12
Journal of Political Science Education (JPSE)	0	0	5	7
Political Science and Politics (PS)	0	0	3	4
Political Science Educator	0	1	2	1
Total	0	6	12	24

These outlets were selected because most papers and articles dealing with teaching and learning appear in these outlets. *Journal of Political Science Education* (JPSE) and *Political Science and Politics* (PS) are journals, and the *Political Science Educator* is the Political Science Education Section newsletter of the American Political Science Association.

four years in some leading outlets for the scholarship of teaching and learning in our field.[3] Clearly, AI was not on the radar of teaching political scientists prior to the advent of ChatGPT 3.5 in the fall of 2022. After that point, the number of papers and articles has doubled each year.

Next, we categorize the papers by theme.[4]

11.4.1 Theme 1: Confronting the Ethical and Equitable use of Generative AI

While not all teaching political scientists are opposed to the classroom use of AI, it is probably fair to say that the vast majority have struggled with questions of ethical and equitable use. The literature on AI reminds us that not all students have the same degree of access to AI tools, and few students come to us equipped with an AI ethical compass in which they are confident (Cantwell-Chavez and Davis, 2025). Some research also explores confronting the ethics of AI explicitly, by designing assignments or class activities that examine, for example, the impact of AI on the Global South or on traditionally underrepresented students (Kehlenbach, 2024).

[3] The counts here exclude considerations of AI as a tool in the broader world of politics, focusing exclusively on its use in the teaching and learning aspects of political science. For comparison, at the 2025 APSA Annual Meeting, there were a total of 105 panels with AI mentioned in the title or abstract of at least one paper.

[4] The list of papers covered in this section appear in Table 11.A. Efforts were made to reduce papers to a single category when feasible. In no case was a single paper assigned to more than two categories, but the fact that some were assigned to two categories explains why the percentages tally above 100 percent. Moreover, in the case of the three APSA Roundtables, there was insufficient information for categorization.

PROOFS

11.4.2 Theme 2: Detecting and Punishing AI Use

Another path of research (14 percent) focused on efforts to identify and/or sanction use of generative AI by students. Taking an approach analogous to plagiarism detection, these studies (which have largely been empirical) have employed techniques ranging from surveys of current practice (Wu and Wu, 2024) to methods for detecting use (Keener, 2025).

However, as scholars have noted, reliable AI detection technologies have not yet been developed and will invariably lag behind AI-generative technologies (Keener, 2025). The risk of false positives is too great to use current versions at scale. A surveillance regime might not even be desirable. On the one hand, it would go a long way to making instructors' lives easier, without worrying about AI-generated integrity violations. On the other hand, it would do nothing in the way of teaching students what technology can and cannot do and would also not address other kinds of cheating. It would, to that extent, solve a real problem with a clunky and imperfect fix.

11.4.3 Theme 3: Department and Syllabus Policies

Some political scientists have engaged with questions of policy: What should my own policy be on AI use? Should my department or university set standards or rules regarding the use of AI? These studies acknowledge both the lack of campus or department-wide policies at most institutions and the need for policy clarity at the classroom level. Some of these studies have gauged student preferences (Cahill and McCabe, 2024), while others have made recommendations for specific class or university policies (Spindel and Ackerman, 2025).

11.4.4 Theme 4: Preventing or Discouraging Use Through Design

Many articles have focused on redesigning course assignments and assessments to make inappropriate uses of generative AI less likely, less desirable, or less effective. The goal in these studies is not a total ban on AI use but an emphasis on maintaining student critical thinking. These studies have included discussions of refocusing evaluation on work that students perform in the classroom (Ringmar, 2024) and tests of assignment wording to encourage student reflection.

11.4.5 Theme 5: Faculty Development and Incorporation

By far, the most common category consisted of research that addressed the value of instructors using AI features: to better design our courses, present class materials, and develop assignments. About 40 percent of the articles and papers we reviewed approached AI optimistically – as a potential ally when employed in the right context. For example, one article considered the specific ways in which AI tools composed Model United Nations papers more or less

PROOFS

effectively than students and how those differences could be used as a teaching tool (De Maio et al., 2025). Another explored using AI tools to improve teaching and learning about literature reviews (Tuncel, 2025).

11.4.6 Theme 6: Career Preparation and Software Use

Several publications and projects (17 percent) focused on the value of harnessing AI to prepare our students for the job market and future careers. Banking on the marketability of an AI-savvy resume, these papers focus on how to teach a particular software (Bergbower and Valdes, 2024) or skill set (Best and Mallinson, 2024).

So, what conclusions can be drawn from this brief metanalysis? Three key takeaways emerge. First, there are concerns about the ethical use of AI and cautions have been raised regarding the embrace of it by political scientists. However, there has also been an increase in exploring AI use in a wide variety of contexts. It's not all about plagiarism or wholesale adoption; political scientists want to carefully consider all the implications of this emerging technology for their classrooms. Finally, political scientists have been trying to devise ways to constructively use this new technology. Rather than writing it off as inappropriate and unethical, or adopting it uncritically to be on the cutting edge, political scientists have begun to approach AI as any other new pedagogical tool. There are also ways in which political scientists have begun to explore the benefits of teaching AI skills to our students for the purposes of career preparation.

In the following, we explore some of the ways in which political science teachers have confronted or engaged with AI.

11.5 THE ETHICS OF AI USE IN COURSES

Before we can consider the role of AI in the classroom, we should begin by asking if AI tools should even be allowed into the classroom space at all. A growing number of higher education instructors are vocalizing concerns with the ethics of permitting and encouraging AI usage in political science classrooms. Kehlenbach (2024) argues that, as political science educators, we ought to examine our individual and collective values with respect to the social, infrastructural, and environmental implications of using generative AI in classrooms.

Indeed, as faculty think about how to address AI usage in the classroom, there are some things that faculty should consider that may affect their decision on whether and how to use AI. It is important to note at the outset that all the major AI forms and structures are privately owned technological apparati designed to create profit for their overarching companies. Even the not-for-profit status of some of these enterprises, like sections of Open AI, serve only to disguise or obfuscate the profit-focused private elements of their business

PROOFS

practices (Hu et al., 2024). In this way, we should always be aware of why a company wants to enter the educational space and how their entry into this space directly impacts their profit-seeking motivations. For example, much has been said about Canvas partnering with OpenAI to integrate ChatGPT into its learning management system. Much less has been said about how this move came only after Canvas' parent company, KKR, invested heavily in AI infrastructure. This is not to say that these tools are undesirable, as that is a personal choice that each instructor must make; but it is important we not think of these technological tools as simple public goods.

There are also data privacy concerns that have implications for teaching and learning. Generative AI models are trained on data that is often obtained without the author or creator's permission, raising copyright and ownership concerns. Moreover, on the user side, the United States has weak data privacy laws for consumers. Currently, user interactions with generative AI fall under the Third-Party Doctrine, which states that any information volunteered online or stored on a company server carries no lawful expectation of privacy. In other words, the government can access private citizens' interactions with ChatGPT by a subpoena (Gilman, 2025). Within the political science discipline, this is concerning as our students may be writing papers on right-wing populism, political violence, or any number of politically charged topics. As political science educators, we should, at minimum, be aware of the lack of legal protections for our students, who may be encouraged in our courses to use AI for engaging with politically divisive topics.

Another consideration that should be at the front of mind for an educator is to resist the deterministic urge to say that AI is "inevitable" or the technology of the future, and we, therefore, need to prepare our students for an era of AI. This claim is largely speculative, and there is good reason to suspect that the economy surrounding generative AI will experience a contraction. MIT recently released a report claiming that 95 percent of companies that adopt AI get no return on their investment, and only 5 percent of companies that explore or pilot AI programs end up adopting them long-term (Challapally et al., 2025). This might be due to so-called Workslop where AI produces content that "masquerades as good work, but lacks the substance to advance a given task," a comment that will, no doubt, ring true to any educator who has had to read a pile of AI-generated essays (Niederhoffer et al., 2025). This combined with chasing educational trends is not a surefire way to ensure student employment (Grove, 2024). Assuming that AI must be a part of the curriculum because it is undoubtedly the way of the future is to commit oneself to a future outcome that is, at best, uncertain.

In addition, one should be aware of the social and environmental impacts of AI before we work to implement it in our classrooms. The massive amounts of continual data collection that is involved in developing and training AI systems, both predictive and generative, raise large concerns about both data privacy and the political and social power involved in collecting data for commercial

PROOFS

use – what some call "datafication" (Couldry and Mejias, 2019; Kehlenbach, 2022; Koopman, 2019; Zuboff, 2019). These outputs of AI, both predictive and generative, have social impacts beyond the issues raised by data collection and management. Because the initial training data used by AI systems is often obtained through convenience or tradition, persistent social biases remain (Crawford, 2021). Kate Crawford traces how AI systems scrape data from Reddit, as well as the corpus of Enron emails that were digitized as part of the legal fallout of Enron's 2001 bankruptcy. The outputs of the systems trained on this data replicate the existing forms of injustice and inequality found within. This is certainly a risk, especially if we unthinkingly embrace AI without serious consideration of the consequences. A note by Rachel Gillum, Gregory Leslie, and Cara Wong in Chapter 5 states that AI helps perpetuate racial inequality because it does not "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."

This is most prevalent within the predictive AI systems used to make public policy decisions, as Virginia Eubanks describes (Eubanks, 2018). However, it is also present in nearly all algorithmic and automated systems, from copy machines to social networks, legal systems, law enforcement, and, perhaps most concerningly for our purposes, educational technology (Benjamin, 2019; Cohen, 2019; Jefferson, 2020; Mirrlees and Alvi, 2020; Moore et al., 2021; Noble, 2018).

Beyond these social and political impacts, the environmental and climate impacts of AI systems are only beginning to be studied. This is partially because the corporations that operate these AI systems have not been forthcoming with accounting of their environmental impact, leading scholars to adopt creative methodologies for attempting to trace the environmental impact of the use of AI. Researchers at the University of California, Riverside, were able to calculate that ChatGPT emitted 502 tons of CO_2 into the atmosphere in 2022. This amount is equivalent to driving 112 gasoline-powered cars for an entire year (Luccioni et al., 2022; Strubell et al., 2019). This also does not account for the electricity usage of AI, which has spiked so high that US consumers are seeing increased energy bills and even calls to reduce personal energy and water usage so as not to overload a network stressed by AI computations (Redman, 2025; Steele et al., 2025). These AI systems, and the datacenters that support them, also raise concerns about environmental degradation, increased fossil fuel use, and increased water consumption (Brook and Sainz, 2024; Wells, 2025). Political science teachers will need to consider these issues when deciding whether to use AI in the classroom.

11.6 ASSIGNMENT DESIGN AND AI

Others have explored how to design assignments to be either AI immune or incorporate the use of AI to promote its responsible and ethical use. Indeed, the

PROOFS

Box 11.1 AI in coursework

AI tools (e.g., ChatGPT) may be used under certain conditions, but use must be transparent and responsible. Instructors reserve the right to require additional process documentation if questions about AI use arise.

- Permitted Use: Brainstorming, outlining, or drafting assistance, if explicitly allowed. Students must submit their AI prompts and outputs with the final assignment and reflect on how they evaluated and revised the AI content.
- Prohibited Use: Submitting AI work without acknowledgment; using AI when independent writing or reasoning is the stated outcome; fabricating sources or evidence.
- Grading: Assignments will be evaluated on both content and critical engagement with AI, including ethical use and reflection. Undisclosed or inappropriate use will be treated as a violation of academic integrity.

rise of generative AI introduces new concerns about assignment design in the political science classroom. On the one hand, instructors face practical challenges in distinguishing student-written work from AI-assisted outputs; on the other, AI presents opportunities to streamline and potentially improve grading practices.

Another challenge to assessing student work in the era of AI is interpretative. Even when AI use is acknowledged, how should instructors evaluate AI-assisted work? Is assessment based on content knowledge, critical thinking, accuracy in data analysis, or writing proficiency? Each of these criteria requires different approaches to grading. Without clear guidance, faculty risk inconsistency across courses and inequity among students, particularly when expectations about AI use vary implicitly, rather than being made explicit. We do not offer such guidance in this chapter but suggest that faculty members consider the example of a statement regarding usage that has appeared in the syllabi of one of the contributors to this chapter (see Box 11.1).

For example, one way to operationalize this is through clear and transparent course policies. Instructors will want to set explicit course policies on acceptable and unacceptable uses of AI, sharing these guidelines in syllabi and class discussions (Spindel and Ackerman, 2025). If possible, students can be involved in the process, sharing their own experiences and expectations, giving them some agency over technology use and a say in the consequences in the course (Donahoe, 2025). Whatever the policy, instructors should consider requiring students to document their use of AI at each stage of the writing process, reflecting on how it shaped their work (Cantwell-Chavez and Davis, 2025; Wu and Wu, 2024). This could be a standalone technology disclosure form or, even better, as part of an open-ended metacognitive reflection and

PROOFS

self-assessment, provided that process requirements and documentation do not become unnecessarily restrictive or burdensome (Spindel and Ackerman, 2025). Gathering this information over the course of the semester, rather than merely at the end, would allow instructors to make timely adjustments that could mitigate achievement gaps and the digital divide.

Language instructors might consider adapting the text in Box 11.1 to their syllabi if they allow the use of AI in the class (this advice was provided by Jennifer DeMaio, one of the authors of this chapter).

11.7 ASSIGNMENT DESIGN

A common topic in the literature on AI use in the teaching and learning of political science relates to assignment design. There are some important principles for assignment design that are noted in the literature. In his book *Cheating Lessons*, James Lang (2013) found that students cheat most when the learning environment combines three conditions: high stakes, high pressure, and low intrinsic value. When courses minimize those conditions and replace them with frequent practice, meaningful engagement, and clearer pathways to success, cheating drops because students feel less cornered. Instructors can minimize the inappropriate use of AI by making greater use of lower stakes assessments, creating an environment where failure is possible and even rewarded and where students are made to understand how what they are learning and expected to do is relevant to their lives and professional goals.

An idea suggested by Donahoe (2025) referring to assignment design involves faculty giving students options in terms of their preferred learning pathway. In this model, students could choose between "AI-free" and "AI-integrated" coursework. The latter option would entail the faculty and students working together to set parameters for the use of AI, as well as penalties for unauthorized usage (Donahoe, 2025).

To that end, faculty should avoid "disposable assignments," which is coursework that has no meaning beyond the course. Instead, we should aim to create "renewable assignments" – that is, important and valuable public-facing projects, such as policy advocacy campaigns, multimedia exhibits, or podcasts, which require original synthesis, critical thought, and real-world implications (Wiley, 2016).

Time permitting, instructors could dedicate more class time to higher order skills and working on assignments in class. This would give faculty a glimpse into how students are using AI tools and give them an opportunity for intervention and instruction. It would also shift the emphasis from product and output to process and mastery (Michels, 2024).

In the following section, we consider ways in which the political science instructor may adapt existing assignments and introduce such assignments in the classroom to promote the skills identified in the 2024 task force report. This generally involves either thinking about how to redesign or reconfigure

PROOFS

existing assignments or introduce new types of assignments into the classroom. These are designed to make the assignments AI-proof or to encourage the responsible and ethical use of AI and ensure students learn the material they are being taught. We offer examples of both suggestions regarding assignment design and redesign. These are *not* meant to be a collectively exhaustive list but may be useful to political science instructors as they consider designing assignments for their courses.

11.7.1 Adapting Assignments to Promote Student Skills

There are many potential ways to redesign existing assignments to address the use of AI and to promote critical thinking, problem-solving, and communication skills in political science students. (For explicit goals, see descriptions in the 2024 APSA task force report.) One simple solution involves making greater use of written assessments with oral exams or in-class reflections, making it harder for a student to rely solely on AI for assessed work. Instructors might also consider offering alternative, creative assignments, for example, memes with explanatory essays, policy briefs, or political ads – alternatives that require critical synthesis, creativity, and personal insight, which is difficult for AI to fake convincingly (Wu and Wu, 2024). Faculty could also require multimodal submissions – such as a video explanation to go along with written work – so students will be required to understand the work that they submit (Michels, 2024).

Perhaps most importantly, scaffolded assignments (or assignments which build upon each other in a sequence) help students to understand the steps required for satisfactory work. This has become even more essential since AI has been introduced into learning. Major assignments need multiple stages: topic proposals, annotated bibliographies, rough drafts, and revisions, so students build work iteratively and receive feedback at each step (Ardoin and Hicks, 2024). This will allow instructors to make sure that students are using AI in appropriate ways and, more importantly, it would permit formative assessments and allow instructors to offer guidance along the way.

Time permitting, instructors will want to dedicate time for in-class writing and assessments, which allows for more immediate feedback and can reduce the temptation for students to use generative writing tools (Ardoin and Hicks, 2024). Relatedly, instructors can expand the use of formal or informal group or individual presentations, panel discussions, or debates to ensure a deeper understanding and discourage surface-level engagement, which is the product of much AI-generated text (Michels, 2024).

Operating at the bottom of Bloom's taxonomy of learning – that is, requiring students to provide only a description, explanation, summary, and so on – creates an incentive for students to find shortcuts. Instead, instructors should design assignments that require higher order thinking and writing, such as the application of concepts, especially to time-sensitive or local events, and case

PROOFS

studies, which are less likely to have been included in AI training data. In particular, using the open-end project-based learning assignments advocated by Chad Raymond (2024) are good examples that require the use of these skills. This could also mean regularly updating assignment topics to focus on current events or recent scholarship. Including students to make references to their own life and experience can also help in that regard. The same is true for active learning, such as problem-solving exercises, open-ended inquiry, and supervised collaborations, especially if they require empirical research or primary data analysis (Ardoin and Hicks, 2024). This may not entirely circumvent more advanced AI models, but it will make students more critical consumers of what AI produces and help them understand how it might be utilized as a useful tool.

The literature on AI and political science teaching also suggests developing new types of assignments to promote student critical thinking, problem-solving, and communication skills (and others as well). For instance, another type of assignment, which can be effective in limited ways, involves integrating AI-assisted brainstorming or outlining as an in-class activity but require students to annotate, critique, or revise output from AI tools to foster critical analysis and awareness of AI limitations (Cantwell-Chavez and Davis, 2025). A similar assignment would include genre translation tasks, which would require students to translate academic writing into public formats with AI assistance, followed by reflection on the adaptations involved. To minimize the temptation and opportunity to use AI-generated text in inappropriate ways, students could work in groups. Relatedly, class activities could include "body-doubling" sessions, where students work on assignments alongside one another, using AI for support rather than replacement, enhancing focus and productivity (Donahoe, 2025). Apart from regulating AI use, this practice can also help students who are prone to getting distracted.

Some scholars have argued that merely adapting existing common assignments are not enough to meet the challenge of AI. Indeed, Chad Raymond (2024) argues that "minor tweaks to assignments and exams will do little to counter the fact that chatbots can distill thousands of pages of scholarship down to a few paragraphs in seconds and may have Ph.D.-level intelligence in only a few years." Instead, he argues that we need to move away from teaching the "what" students should know (involving a student passively ingesting a heavily curated body of information for later regurgitation) to teaching students how to be builders of new knowledge through creative problem-solving.

To do this, Raymond argues that political science instructors should engage with project-based learning that stimulates students' intrinsic curiosity while utilizing AI to achieve this goal. In practice, this means having students select an unresolved real-world problem of interest that is relevant to the course's subject and then make them responsible for inventing a solution to the problem they have chosen. Projects with such indeterminate outcomes elicit emotional investment and original thinking from students, making AI use less attractive.

PROOFS

Indeed, although AI systems excel at retrieving, analyzing, and synthesizing information, they are far less able to form and evaluate options for dealing with complex political problems. That is why Raymond offers assignments so that students grapple with unstructured problems that have no single correct answer. In a comparative politics class that he offers, he asks students to assume the role of political-risk consultants and, within twenty minutes, identify which of four African states is the most suitable location for a German company's new manufacturing plant. Students can use AI to explore various options but must justify their reasoning and reflect on what they learned about problem-solving as part of the assignment.

Further, Raymond (2024) contends that an important additional advantage of using problem-solving and project-oriented learning is that such projects are conducive to making learning a team sport. In most organizations, people work in groups and productive collaboration is critical to success. Good ideas can arise from conversations with colleagues, and people frequently learn tremendous amounts by teaching others. Working on a shared goal also fosters a sense of community. Artificial intelligence cannot yet replicate the benefits that come from these human interactions.

Another assignment that uses AI explicitly to help build civil discourse and communication skills is illustrated by the work of Christopher Bail (Armstead, 2024). In a 2024 study "Bridging Political Divides with Artificial Intelligence," he argued that LLMs, such as ChatGPT, can be used for conflict mediation and make political conversations more effective and civil (Armstead, 2024). His premise is that the current political atmosphere in the United States is much like a situation facing a marriage counselor where "people talk past each other." In his study, a group consented to have a conversation with a partner who had a different political view than themselves. The discussion was about gun control, and the left-leaning participant described guns as a "stain on democracy," while the politically right-leaning person would respond that "guns are protecting democracy." The debate would then end with one side ridiculing the other as "incredibly naïve." Bail then provided participants in the first trial the opportunity to use AI chat assistant technology that offered alternatives to participants to not use responses involving insults. The chat assistant gave the user a choice to rephrase their message to make the conversation more productive. The use of the AI chat assistant led to better quality civil discourse and increased democratic reciprocity but would not change participants' minds on policy issues. However, Bail found that by using ChatGPT, the users were able to take the perspective of their partner. By stepping into their adversary's shoes, they better understood the other's point of view. As a result, participants who used and received chat assistant messaging were more inclined to have future tough conversations as a classroom activity.

Artificial intelligence use has also been cited by some scholars as a means to help improve student writing. In its simplest form, AI can be used as a grammar and spell checker to help students catch basic errors and improve

PROOFS

technical writing quality (Wu and Wu, 2024). This is especially good for non-native English speakers, who can use AI to suggest edits for structure or style, making their writing more accessible and less of a barrier to learning. In addition to helping them with the task at hand, it can also help them to learn the language quicker, with immediate feedback and personalized help. The same is true for lesser-prepared and first-generation students.

Students could also be encouraged to use AI to rewrite or paraphrase drafts for clarity, conciseness, or tone before finalizing their submissions. This could save instructors countless hours of elementary writing feedback, which could be better spent on higher-level instruction (Michels, 2024). In this way, faculty shift from content delivery to mentor and coach, helping students create, critique, and challenge knowledge, rather than passively receive it.

Instructors could also employ AI to help generate genre-specific academic, professional, and popular forms of writing, such as policy briefs, abstracts, and op-eds, so students can analyze and adapt their style accordingly (Spindel and Ackerman, 2025). Alternatively, feedback from AI could be used for micro-writing challenges, where students revise short, unfocused paragraphs for specific style goals.

Assignments could focus on the annotation and evaluation of sources, emphasizing proper citation practices and awareness of AI's propensity for hallucinated references (Wu and Wu, 2024; Dilling and Owen, 2025). Artificial intelligence could also be used for brainstorming sessions, where the output is used for analytical exercises and the development of original ideas afterward. Similarly, instructors can use the output to teach "AI literacy" and help students to recognize and navigate biases, inaccuracies, and ethical pitfalls in AI outputs (De Maio et al., 2025). And in "reverse outlining" exercises, students could have AI extract and organize main points to improve argument flow and coherence.

PROOFS

11.7.2 Other Applications of AI in Teaching and Learning

Beyond developing the skills outlined in the 2024 Presidential Task Force Report, there are other cited uses of AI in the current literature. These include using AI for the grading and evaluation of student work, employing AI to improve student digital literacy, and teaching AI use for career preparation (which is a focus of the 2024 report). In Section 11.7.2.1, we discuss the literature that has touched upon these three themes.

11.7.2.1 Using AI for Grading and Evaluation

Another body of literature that is relevant for instructors considering the use of AI is its use in grading and evaluation of student work. As we noted, there are concerns about the use of detection software to detect AI use (which are forms of Predictive AI), and these concerns merit serious consideration. Some scholars argue that AI-assisted feedback, when used transparently and as a supplement

to instructor judgment, can support more equitable learning environments by providing timely, consistent feedback and helping students develop research literacy and communication skills. Platforms such as Packback (https://packback.co/) incorporate predictive AI-powered grading assistants that provide rubric-based scoring focused primarily on technical and mechanical elements of writing. These digital assistants reduce grading loads, enabling instructors to devote more of their feedback to higher-order concerns, including argument quality, substantive content, and student mentoring (Selwyn, 2021). The use of AI-generated feedback on student writing, however, "raises the possibility that AIs are now evaluating AI-generated papers, reducing the entire academic exercise to a conversation between two robots – or perhaps even just one" (Walsh, 2025).

To balance the challenges and opportunities to grading introduced by AI, faculty could implement several strategies. Grading practices could align with clearly defined policies, ensuring equity and clarity for students. In addition, rubrics could be expanded to include criteria for reflection, ethical use, and discipline-specific application, thereby allowing for transparent, purposeful integration of AI. If instructors want to use AI detection tools, they could improve their reliability by requiring process evidence, such as drafts, revisions, and/or reflections.

These strategies offer a practical approach to thinking about grading in the age of AI, but the greatest opportunity for improving how we evaluate student work may lie in critical examination, both by individual instructors and by institutions, on the ways grading practices can reproduce inequality. Research in higher education has long shown that grading often privileges dominant linguistic styles, implicit norms of academic professionalism, and markers of cultural capital that disadvantage first-generation students, multilingual learners, and students with learning differences (Wargo and Anderson, 2024; Inoue, 2015; Elbow, 1993; Bourdieu, 1986). Artificial intelligence tools may assist instructors in identifying patterns in grading outcomes and feedback that may otherwise go unnoticed, such as systematic differences across student groups, assignment formats, or modes of participation (Feldman, 2019). Artificial intelligence–assisted review can also prompt instructors to examine the language of their comments. It can help them identify bias, recalibrate rubrics to emphasize learning objectives (rather than surface-level polish), and design assessments that align with principles of universal design for learning (Center for Applied Special Technology, 2018). At the same time, these tools are not neutral: AI systems are themselves shaped by existing power structures and data biases and must be deployed cautiously to avoid reinforcing the inequities they seek to address (Benjamin, 2019; Noble, 2018). Used as a reflective and diagnostic aid rather than a substitute for human judgment, AI has the potential to support more equitable, intentional, and pedagogically grounded approaches to grading in political science classrooms.

Reframing the discussion of the impact of AI on grading from policing to a pedagogical opportunity requires a shift in how instructors conceptualize

PROOFS

grading and evaluation, particularly with respect to equity and bias. There remain very real challenges to grading presented by AI, and political science instructors should be careful in their decision to use these tools. But these tools may be helpful to reduce long-standing inequities in evaluation. Rather than focusing on the detection of AI use, instructors can design assignments and grading criteria that emphasize learning processes, revision, and metacognitive engagement, elements that have been shown to reduce bias.

11.7.2.2 Using AI to Promote Digital/Information Literacy

Although not explicitly part of the skills recommendations of the 2024 presidential task force (APSA, 2024), information literacy is clearly an extremely important student skill in the era of AI. Information literacy involves "the set of skills needed to find, retrieve, analyze, and use information" (Marfleet, 2005). Instructors will need to help students hone their information and media literacy skills with in-class exercises on the proper use of AI (Topal and Shargh, 2023). It is even more reason why information literacy cannot simply be seen as just an extension of the library.

The introduction of AI in the political science classroom raises important questions about how to assess student learning around digital literacy. While information literacy has already been included as a student learning outcome (SLO) in many programs (Lowe and Piper, 2021; Sparks et al., 2016; Saunders, 2011), AI tools expand both the scope of what constitutes information literacy and present challenges with how to measure it. As critical engagement with AI-generated information becomes a core competency, and one likely to be essential for students' professional trajectories, it is increasingly important to develop assessment models that capture AI literacy. Departments are now tasked with designing meaningful assessments and rethinking digital information literacy SLOs alongside rapidly evolving technology.

While the challenges of assessing digital literacy in the age of AI are considerable, there are opportunities to think about assessment in productive ways. Departments could explicitly articulate digital literacy around AI as an SLO, develop rubrics that measure it across courses (see sample rubric below), and incorporate AI-enhanced assignments to gather evidence of student learning over time (Banta and Palomba, 2015). But to integrate digital literacy around AI into assessment, programs must begin by considering definitions, evidence collection, and closing the loop through program improvement. An example of how one might measure digital literacy has been offered by a recent article by Andrew Guess and Kevin Munger (2023). They provide a very useful measure that could be used to gauge student progress in digital literacy skills.

One of the most fundamental challenges comes with defining what should be assessed. Digital information literacy in political science has traditionally been understood to encompass technical skills, such as navigating databases,

PROOFS

coding, or evaluating online sources. This definition must now be broadened to include the ability to critically interpret algorithmic biases and assess the credibility of AI-generated outputs. Beyond the need to evaluate a new set of skills, ethical considerations also take on new dimensions as students navigate questions of academic integrity and transparency in their use of AI (Cahill and McCabe, 2024). These challenges in redefining digital information literacy raise the subsequent question of how programs can meaningfully assess the outcomes they identify.

As they articulate definitions and develop SLOs, programs must think about designing assignments and instruments that generate meaningful evidence of student learning. Generative AI complicates the interpretation of such evidence: If students rely on AI to write, generate ideas, or analyze data, how can programs accurately evaluate their learning? Equity concerns also emerge, as students with greater familiarity or access to AI tools may appear to perform at higher levels in ways not necessarily reflective of deeper learning (Selwyn, 2021). One approach is to require students to interact directly with AI tools under clearly defined expectations for annotation, reflection, and critique (Alexander, 2023). For instance, students might produce an AI-assisted policy memo or legal brief and then evaluate its strengths, limitations, and embedded biases. Such tasks enable programs to assess both technical proficiency and higher-order analytical skills simultaneously. In addition, portfolios, annotated outputs, and reflective essays can provide richer, longitudinal evidence of students' developing digital literacy around AI (Banta and Palomba, 2015). This evidence provides the basis for program-level reflection and improvement.

With assessment instruments in hand, departments can close the loop by interpreting results and making programmatic adjustments that respond directly to demonstrated areas of need. Systematic analysis of student learning data allows departments to move beyond anecdotal impressions and, instead, ground curricular decisions in documented patterns of performance. By developing rubrics to track student growth across multiple courses, departments could identify where additional instruction or scaffolding is needed, as well as where students are successfully demonstrating emerging competencies (Alexander, 2023). Such longitudinal assessment tools also make it possible to differentiate between issues rooted in course design and those stemming from broader program-level gaps. Assessment data can additionally inform faculty development, ensuring that instructors receive targeted training in the pedagogical and ethical use of AI and are equipped to guide students through its technical, methodological, and normative complexities (Williamson, 2016). In this way, assessment not only measures student learning but also functions as a mechanism for strengthening instructional capacity and aligning departmental practices with rapidly evolving expectations around AI literacy. Table 11.2 offers an example of a rubric that might be used to assess digital literacy with AI tools.

PROOFS

TABLE 11.2 *Sample rubric-assessing digital literacy with AI tools in political science*

Student Learning Outcome: Students will demonstrate advanced digital literacy with AI tools by analyzing AI-generated information for accuracy and bias, evaluating its relevance and ethical implications for academic inquiry, and integrating selected AI outputs to create original, well-supported scholarly work.				
Criteria	**Exemplary (4)**	**Proficient (3)**	**Developing (2)**	**Beginning (1)**
Technical Proficiency	Effectively integrates AI tools into political science tasks (e.g., research, data analysis, writing). Demonstrates fluency with multiple platforms.	Uses AI tools competently with minimal guidance. Selects appropriate tool for the task.	Demonstrates partial or inconsistent ability to use AI tools. May rely on trial-and-error.	Struggles to use AI tools even with assistance. Misapplies or avoids technology.
Critical Evaluation	Consistently identifies strengths, limitations, and biases in AI outputs. Provides well-reasoned analysis with supporting evidence.	Identifies most strengths and weaknesses of AI outputs; analysis is generally sound.	Demonstrates limited or surface-level critique of AI outputs; may overlook major issues.	Accepts AI outputs uncritically; little or no evidence of evaluation.
Ethical and Responsible Use	Consistently applies ethical standards: transparency in attribution, awareness of bias, adherence to academic integrity. Reflects on implications for political science.	Usually applies ethical standards and acknowledges limits of AI. Minor lapses in transparency.	Inconsistent application of ethical standards; limited reflection on implications.	Rarely or never applies ethical standards; evidence of inappropriate or dishonest use.

(*continued*)

PROOFS

TABLE 11.2 (*continued*)

Student Learning Outcome: Students will demonstrate advanced digital literacy with AI tools by analyzing AI-generated information for accuracy and bias, evaluating its relevance and ethical implications for academic inquiry, and integrating selected AI outputs to create original, well-supported scholarly work.				
Criteria	**Exemplary (4)**	**Proficient (3)**	**Developing (2)**	**Beginning (1)**
Application to Political Science	Applies AI tools to discipline-specific problems with sophistication (e.g., policy memo, data interpretation, theory critique). Demonstrates disciplinary insight.	Applies AI tools to political science tasks appropriately and with some insight.	Attempts to apply AI to political science but with limited success or superficial understanding.	Minimal or no evidence of applying AI to political science contexts.
Reflection	Provides thoughtful, detailed reflection on the learning process, decisions, and implications for future use.	Provides adequate reflection on process and decisions; some discussion of implications.	Reflection is present but minimal, vague, or disconnected from practice.	No meaningful reflection; unable to articulate process or implications.

PROOFS

11.7.2.3 Using AI for Career Preparation

Finally, political science scholars are also thinking in terms of how AI might be used to better prepare students for the current employment market (Moulton, 2024). Although this literature is generally less developed than other aspects of the use of AI in teaching and learning political science, some scholars have argued that for today's college graduates, the twenty-first century workforce landscape requires more than just a conceptual grasp of institutional hierarchies, power, or systemic dynamics. It requires the technical dexterity to measure, map, and mobilize them.

The transition from the classroom to the workforce is caught in a tug-of-war between preparing students for a tradition-bound academic track and meeting the relentless demands of an automated workforce – a struggle that increasingly frames higher education as being out of touch with the practical realities of the silicon age. Despite existing challenges surrounding the

discipline, global volatility and the rapid rise of automated intelligence have created opportunities for curricula innovation within the discipline. Political science programs may need to rethink ways in which to innovate the current curricula (Moulton, 2024).

In a technocratic era, curricula innovation should move beyond the study of civic leadership to the active cultivation of it. Using AI-generated simulations and training, political science programs can integrate into the curricula training in crisis communication, consensus-building, and restorative justice. These trainings will prepare graduates capable of mediating complex public disputes, managing multi-stakeholder coalitions, and navigating the ethical challenges of governance. Programs can produce graduates who possess not just theoretical rigor but also the "soft skills" that are highly required to thrive in today's workforce landscape. Such curricula innovation transforms the model of civic engagement from a mere volunteer activity into a professional proficiency, enabling graduates to spur community-based policy initiatives, lead legislative roles, and drive nonprofit works (APSA, 2024).

To master and navigate the complexities of modern governance, students need a "flight simulator" for the study of politics. Integration of games and simulations into the curricula would enable students to apply abstract theories like Game Theory or Just War Theory in real-time. By gamifying the curricula, programs can cultivate rapid response decision-making skills and analytical strategic foresight among students; skills that are highly required in corporate strategy, conflict management, and intelligence roles (Moulton, 2024).

In a silicon-based workforce that is increasingly reshaped by the automation of tasks, the ability to perform cross-domain synthesis has become a valuable professional differentiator. While generative AI tools can summarize, synthesize, or distill billions of the world's data, information, and knowledge in a matter of seconds, these tools remain a "black box" that requires human interpretation of the normative and systemic consequences of their own outputs. As Nick Bostrom (2014) articulated in this work on the "value alignment problem," these machines do not understand human values. As such, political science is one of the few disciplines that train students to "unbox" these complex human values and systems, ensuring that institutional outputs are capable of aligning with democratic and ethical standards (McGrath and Diaz, 2021).

Integrating these curricula innovations would prepare graduates to look beyond an algorithmic prediction and analyze the intersectional impact of variables – understanding, for example, how a shift in domestic trade protectionism policies might ripple through the global supply chain and logistics, or providing intelligence risk assessments to corporations when political variables like populism impact both the domestic and global economy. While AI can calculate a correlation, a trained "political practitioner" can interpret the "why" behind correlation and advance a strategic "how" for the present and future.

PROOFS

11.8 STANDARDS FOR AI USE IN THE CLASSROOM?

As we noted previously, the appearance of AI and its use by students has raised serious challenges to the implementation of the recommendations made by the 2024 APSA presidential task force report on rethinking political science education. Artificial intelligence use presents serious challenges to the development of critical thinking; problem-solving; and analytical, communicative, and civic skills. In this chapter, which is not meant to be prescriptive, we have outlined how our discipline has reacted to the advent of AI, discussed the current ethical debate over the use of AI in the political science classroom, and illustrated the various ways in which political science instructors have engaged with AI in the classroom. However, it is important to note the technology driving the development of AI is rapidly changing, and that will undoubtedly present even more challenges to political science instructors as they try to teach students skills for both democratic citizenship and the future workforce.

Although we have avoided being prescriptive in this chapter, there are some things we believe our association should consider as we navigate an uncertain future. First, rather than assume AI use in the classroom is inevitable and that we should rush to embrace this new technology, it is important to continue consideration of the ethics and the social and environmental costs of using AI. It is important that the concerns brought up in this report are communicated to political science faculty so that they may make informed decisions on whether to use AI in the first place.

Second, for those colleagues who are interested in using AI in their classrooms (and that appears to be a fairly substantial number after reviewing survey results and reviewing the existing literature), this chapter has tried to highlight some of the things colleagues in the discipline are doing. American Political Science Association has already provided resources on ethical and responsible use of AI in the classroom on the platform APSA Educate. But this will need to be significantly expanded over the coming years, both in the number and variety of topics covered (including AI use and grading, AI use and digital literacy, and AI and the development of career skills).

Third and perhaps most important, political science instructors need some guidance from APSA on what constitutes acceptable use of AI in our classrooms. As we noted earlier, instructors are pretty much on their own in developing policies on AI use. We have suggested, based on the literature, some guidelines for instructors for policies that might be adapted for course syllabi. However, it would be preferable for APSA to provide guidance with regard to AI use in the classroom. We are aware that some efforts are being made in other political science organizations (most notably the International Political Science Association and the International Studies Association) to develop guidance on acceptable AI use in research and publication.[5] Why not develop similar

[5] John Ishiyama, one of the coauthors of this chapter, is currently involved in the IPSA effort to develop such guidelines.

PROOFS

guidance for political science instructors? Perhaps a first step is to convene, via the APSA department services committee, an initiative to bring chairs together at the annual meeting. Through a conference within a conference dedicated to sharing various AI policies that political science departments have adopted, we could fashion a working group that could help develop an APSA-wide set of standards for acceptable use in the classroom. With the organizational weight of APSA behind such an initiative, it will raise the likelihood that departments across the country would adopt such standards, and this would promote a unified disciplinary effort to confront the challenges of AI for teaching and learning. In any case, APSA must confront the challenges presented by AI usage to student learning goals that the association adopted in the 2024 task force report (APSA, 2024).

APPENDIX

TABLE 11.A *AI and teaching papers from APSA annual meetings (including TLC at APSA)*

Year	Author	Paper
2023	(roundtable)	Breaking News: Generative AI and the Future of Political Science
2023	(roundtable)	Teaching in the Time of ChatGPT
2023	Nicole Wu, Patrick Y. Wu	Generative Large Language Models and Political Science Education
2023	Jacob Wobig	The Write Stuff: Designing Undergraduate Writing Assignments in the Age of AI
2023	Leila Farah, Spyridon Kotsovilis, Martin Revermann	Using Ancient and Modern Theatre in the Classroom to Teach Political Science
2024	Kimberley MacVaugh, Rachel Olsen (workshop)	Teaching Information Literacy: Reflecting on Research in the Age of AI
2024	(roundtable)	Transforming Lives through Teaching: The Impact of Community College Faculty on Students
2025	Kimberly MacVaugh	Reimagining Information Literacy Skills in the Age of AI
2025	Blair Thomas	Putting the AI in Political Science: How Instructors Are Navigating a New Classroom
2025	Jacob Wobig	Building Scaffolds, Not Shortcuts: AI in Political Science Writing Assignments
2025	Janet Marie Lawler	Teaching Intentionally with AI: Transparency, Authenticity, and Self-Generation

(*continued*)

PROOFS

TABLE 11.A (*continued*)

Year	Author	Paper
2025	Zora Mihaley	From Chalkboards to Chatbots: Political Science Pedagogy in the Age of AI
2025	Kris Aaron Beck	Gen-AI and Student Learning Outcomes
2025	Andrew Owen	Artificial Intelligence, Academic Integrity, and Writing in Political Science
2025	Kevin Cope	Grading Machines: Can AI Exam-Grading Outperform Professors?
2025	Charles C. Turner, Audrey Thao, and Ginny Adorador	Generative AI, Academic Integrity, and the American Government Classroom
2025	Mike Cowan	Ctrl+Alt+Regulate: Assessing College Students' Preferences for AI Governance
2025	Dror Kris Markus, Fabrizio Gilardi	Interdisciplinarity and Societal Concerns in AI Research
2025	Taekyong Kim and Young Jee Kim	Insights from School-Wide Deliberation at a California Middle and High School

BIBLIOGRAPHY

Alqahtani, N., and Z. Wafula. 2025. "Artificial Intelligence Integration: Pedagogical Strategies and Policies at Leading Universities." *Innovative Higher Education*, 50(2): 665–684.

APSA. 2024. "Presidential Task Force." *Rethinking Political Science Education*. March. https://tinyurl.com/57wydh3v

Alexander, B. 2023. *Universities on Fire: Higher Education in the Climate Crisis*. Johns Hopkins University Press.

Ardoin, P. J., and W. D. Hicks. 2024. "Fear and Loathing: ChatGPT in the Political Science Classroom." *PS: Political Science and Politics*, 57(4): 583–593. https://doi.org/10.1017/S1049096524000131

Armstead, J. 2024. "Political Science Expert Explores Using AI to Talk across the Aisle." https://tinyurl.com/mstbhzyb

Atoyan, V., L. Brosius, N. Movsisyan, S. Ohanyan, and V. Hovyan 2025. "The Role of Artificial Intelligence in Higher Education: The Case of Armenia." *PS: Political Science and Politics*, 58(3): 587–592. https://doi.org/10.1017/S104909652400115X

Banta, T. W., and Catherine A. Palomba. 2015. *Assessment Essentials: Planning, Implementing, and Improving Assessment in Higher Education*. Jossey-Bass.

Baum, J., and Villasenor, J. 2023. "The Politics of AI: ChatGPT and Political Bias." Web Article, Brookings Institution, May, 8. www.brookings.edu/articles/the-politics-of-ai-chatgpt-and-political-bias/

Bender, E. M., and A. Hanna. 2025. *The AI Con: How to Fight Big Tech's Hype and Create the Future We Want*. Harper.

Bender, E. M., and A. Grissom. 2024. "Power Shift: Toward Inclusive Natural Language Processing." In Anne Charity Hudley, Christine Mallinson, and Mary Bucholtz (Eds.), *Inclusion in Linguistics* (pp. 199–221). Oxford University Press.

PROOFS

Benjamin, R. 2019. *Race after Technology: Abolitionist Tools for the New Jim Code*. Polity Press.
Bennett, S., and K. Maton. 2010. "Beyond the 'Digital Natives' Debate: Toward a More Nuanced Understanding of Students' Technology Experiences." *Journal of Computer Assisted Learning*, 26(5): 321–331.
Bergbower, M. L., and J. Valdes. 2024. "Classroom and Student Experiences with Perusall." *Journal of Political Science Education*, 20(3): 496–501. https://doi.org/10.1080/15512169.2024.2331579
Best, E., and D. J. Mallinson. 2024. "Quantitative Political Science Education in the Past and Future." *Journal of Political Science Education*, 20(4): 637–649. https://doi.org/10.1080/15512169.2023.2260034
Blair, A., C. Butcher, T. Bhasin, E. Gordon, M. C. Hallward, A. R. M. McCartney, and S. Usherwood. 2025. "Artificial Intelligence Challenges and Opportunities." *Journal of Political Science Education*, 21(2): 187–191. https://doi.org/10.1080/15512169.2025.2481586
Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.
Bourdieu, P. 1986. "The Forms of Capital." In J. Richardson (Ed.), *Handbook of Theory and Research for the Sociology of Education* (pp. 241–258). Greenwood.
Breuning, M. 2021. "John Ishiyama: Teaching Research and Researching Teaching." *Political Science Today*, 1(3): 18–24.
Brook, J., and A. Sainz. 2024. "Meta to Build $10 Billion AI Data Center in Louisiana as Elon Musk Expands His Tennessee AI Facility." *AP News*. https://tinyurl.com/yekpnar6 (July 3, 2025).
Brubaker, R. 2022. *Hyperconnectivity and Its Discontents*. John Wiley and Sons.
Buchler, J. 2009. "Teaching Quantitative Methodology to the Math Averse." *PS: Political Science and Politics*, 42(3): 527–530.
Cahill, C., and K. McCabe. 2024. "Context Matters: Understanding Student Usage, Skills, and Attitudes Toward AI to Inform Classroom Policies." *PS: Political Science and Politics*, 57(4): 594–601. https://doi.org/10.1017/S1049096524000155
Cantwell-Chavez, D., and J. Davis. 2025. "Navigating Generative AI Tools in the Classroom Through a Lens of Equity and Accessibility." *Journal of Political Science Education*, 21(2): 277–289. https://doi.org/10.1080/15512169.2024.2426153
Carranza Ko, Ñ., and M. Shochet. 2024. "Exploring Open Education Pedagogy in Research Methods Classrooms: Diversifying Methods." *Journal of Political Science Education*, 20(4): 556–571. https://doi.org/10.1080/15512169.2024.2318238
CAST. 2018. *Universal Design for Learning Guidelines Version 2.2*. http://udlguidelines.cast.org
Challapally, A., C. Pease, R. Raskar, and P. Chari. 2025. "State of AI in Business 2025." *MIT NANDA*: 26. www.aigl.blog/state-of-ai-in-business-2025/
Chang, H. 2025. "Social Annotation Platforms: Hypothesis and Perusall." *Journal of Political Science Education*, 21(4): 802–810. https://doi.org/10.1080/15512169.2025.2510623
Cohen, J. E. 2019. *Between Truth and Power*. Oxford University Press.
Collins, T. A., H. G. Knotts, and J. Schiff. 2012. "Career Preparation and the Political Science Major." *PS: Political Science and Politics*, 45(1): 87–92.
Couldry, N., and U. A. Mejias. 2019. *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism*. Stanford University Press.

PROOFS

Crawford, K. 2021. *The Atlas of AI*. Yale University Press.
de la Puente Pacheco, M. A., J. Torres, A. L. Blanco Troncoso, H. J. Guzmán Murillo, and J. X. M. Carrascal. 2025. "Enhancing Critical Thinking and Argumentation Skills in Colombian Undergraduate Diplomacy Students: ChatGPT-Assisted and Traditional Debate Methods." *Journal of Political Science Education*, 21(4): 728–738. https://doi.org/10.1080/15512169.2025.2449936
De Maio, J. L., I. Kabalaki, S. Moshtael, and M. A. Tejax. 2025. "AI versus Students: A Study of the Capability of ChatGPT to Write Model United Nations Position Papers." *PS: Political Science and Politics*, 58(2): 357–362. https://doi.org/10.1017/S1049096524000799
Dilling, M., and L. Owen. 2025. "Designing Politics and IR Assessments in the Era of AI: An Empirical Investigation into ChatGPT's Output across Bloom's Revised Taxonomy." *Journal of Political Science Education*, 21(2): 290–309. https://doi.org/10.1080/15512169.2024.2408767
Donahoe, E. P. 2025 "More on AI and Academic Integrity." *Substack*, May 9. https://emilypittsdonahoe.substack.com/p/more-on-ai-and-academic-integrity
Elbow, P. 1993. "Ranking, Evaluating, and Liking: Sorting Out Three Forms of Judgment." *College English*, 55(2): 187–206. https://doi.org/10.58680/ce19939323
Eubanks, V. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.
Feldman, J. 2019. *Grading for Equity: What It Is, Why It Matters, and How It Can Transform Schools and Classrooms*. Corwin.
Ghaffary, S. (2023, September 21). "Universities Rethink Using AI Writing Detectors to Vet Students' Work." *Bloomberg*. https://tinyurl.com/56beenhf
Gilman, N. 2025. "If You Tell ChatGPT Your Secrets, Will They Be Kept Safe?: Guest Essay." *New York Times* (Online). https://tinyurl.com/ym7jssaj
Grove, J. 2024. "New Study Challenges Conventional Wisdom on Shortage of STEM Graduates." *Inside Higher Ed*. September 23, 2025. https://tinyurl.com/yeyju92z
Guess, A. M., and Munger K. 2023. "Digital Literacy and Online Political Behavior." *Political Science Research and Methods*, 11(1): 110–128.
Henshaw, A. 2024. "Thinking Big? Toward a Pedagogy of Critical Data Studies in Political Science." *Journal of Political Science Education*, 20(4): 623–636. https://doi.org/10.1080/15512169.2024.2306332
Hu, K., and K. Cai. 2024. "Exclusive: OpenAI to Remove Non-Profit Control and Give Sam Altman Equity." *Reuters*. September 23, 2025. https://tinyurl.com/2hcvuz97
Inoue, A. B. 2015. *Antiracist Writing Assessment Ecologies: Teaching and Assessing Writing for a Socially Just Future*. WAC Clearinghouse.
Jefferson, B. J. 2020. *Digitize and Punish: Racial Criminalization in the Digital Age*. University of Minnesota Press.
Juverdeanu, C. 2023. "In Part Usage of AI and Academic Integrity." *Political Science Educator*, 27(1): 34–37. https://tinyurl.com/ms867b4s
Keener, R. 2025. "More A than I: Testing for Large Language Model Plagiarism in Political Science." *PS: Political Science and Politics*, 58(4): 657–663. https://doi.org/10.1017/S1049096525000368
Kehlenbach, E. S. 2022. "The Subatomic Person: A New Ontology of Big Data." *Theory and Event*, 25(4): 23–45.
Kehlenbach, E. S. 2024. "The Impact of Infrastructure: Considerations of Generative AI in the Classroom." *Journal of Political Science Education*. Advance online publication. https://doi.org/10.1080/15512169.2024.2359439

PROOFS

Koopman, C. 2019. *How We Became Our Data: A Genealogy of the Informational Person*. The University of Chicago Press. https://doi.org/10.7208/chicago/9780226626611.001.0001

Lang, J. 2013. *Cheating Lessons: Learning from Academic Dishonesty*. Harvard University Press.

Leslie, R., H. Beckler, D. Campbell, E. Berenstein, R. Makhoul, and T. Merkel. 2025. "We Uncovered the Owners, Locations, and Environmental Impact of America's Data Centers." *Business Insider*. September 23, 2025. https://tinyurl.com/4rss6c4v

Lowe, M. S., and G. F. Piper. 2021. *Envisioning the Framework: A Graphic Guide to Information Literacy*. Publications in Librarianship 77. ScholarWorks, Indiana University–Purdue University Indianapolis.

Luccioni, A. S., S. Viguier, and A.-L. Ligozat. 2022. "Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model." October 5, 2023. http://arxiv.org/abs/2211.02001

Marfleet, B. G. 2005. "Information Literacy and the Undergraduate Research Methods Curriculum." *Journal of Political Science Education*, 1(2): 175–190.

McGrath, K., and J. Diaz. 2021. APSA Graduate Placement Report: Analysis of Political Science Placements. https://tinyurl.com/k3yuwecm

Michels, S. 2024. "Teaching (with) Artificial Intelligence: The Next Twenty Years." *Journal of Political Science Education*, 20(4): 510–521. https://doi.org/10.1080/15512169.2023.2266848

Mirrlees, T., and S. Alvi. 2020. *EdTech Inc.: Selling, Automating and Globalizing Higher Education in the Digital Age*. Routledge.

Moore, S. D. M., B. De Oliveira Jayme, and J. Black. 2021. "Disaster Capitalism, Rampant EdTech Opportunism, and the Advancement of Online Learning in the Era of COVID19." *Critical Education*, 12(2): 24.

Morris, S. M., and J. Stommel. 2017. "A Guide for Resisting Edtech: The Case against Turnitin." *Hybrid Pedagogy*. January 21, 2021. https://hybridpedagogy.org/resisting-edtech/

Moulton, J. F. G. 2024. "What is The Employability Value of a Degree in Politics and International Relations?" *Journal of Political Science Education*, 20(3): 405–421. https://doi.org/10.1080/15512169.2023.2284821

National Association of Colleges and Employers (NACE). 2024. *2024 Student Survey Report: Career Readiness and Competencies*. NACE.

Newfeld, D. 2019. "An Analysis of the Flipped Classroom Approach for "Math-phobic" Students." *Journal for Excellence in Business and Education*, 6(1). www.jebejournal.org/index.php/jebe/article/view/84

Niederhoffer, K., G. R. Kellerman, A. Lee, A. Liebscher, K. Rapuano, and J. T. Hancock. 2025. "AI-Generated 'Workshop' Is Destroying Productivity." *Harvard Business Review*. https://hbr.org/2025/09/ai-generated-workslop-is-destroying-productivity (September 23, 2025).

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

Ntoutsi, E., P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M. E. Vidal, … S. Staab. 2020. "Bias in Data-driven Artificial Intelligence Systems – An Introductory Survey." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3): e1356.

Oldmixon, E. A. 2018. "It Was My Understanding That There Would Be No Math": Using Thematic Cases to Teach Undergraduate Research Methods." *Journal of*

PROOFS

Political Science Education, 14(2): 249–259. https://doi.org/10.1080/15512169.2017.1405356
OpenAI. 2023. *AI Classifier for Indicating AI-written Text: Limitations and Considerations*. OpenAI.
Perrigo, B. 2023. "The $2 Per Hour Workers Who Made ChatGPT Safer." *Time*. October 5, 2023. https://time.com/6247678/openai-chatgpt-kenya-workers/
Raymond, C. 2024. "AI and the Case for Project-Based Teaching." *The Chronicle of Higher Education*. August 27. www.chronicle.com/article/ai-and-the-case-for-project-based-teaching
Redman. 2025. "Environmental Groups Raise Alarm on AI Data Center Use of Energy, Water." *Yahoo News*. September 23, 2025. www.yahoo.com/news/articles/environmental-groups-raise-alarm-ai-205151481.html
Ringmar, E. 2024. "Political Science Exams in the Age of AI." *Political Science Educator*, 28(1): 72–75. https://tinyurl.com/mr299pvc
Rivetti, P., R. Banerjee, and D. O'Mullane. 2025. "Testing ChatGPT in International Relations Classrooms: Potentialities, Limitations, and What's Next." *PS: Political Science and Politics*, 58(1): 69–176. https://doi.org/10.1017/S1049096524000817
Roberts, S. T. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press.
Robertson, J. 2024. "Can Chatgpt and Crowdsourced Forecasting Help Students Think About International Relations? A New Class Assignment." *Political Science Educator*, 28(1): 55–64. https://tinyurl.com/mr299pvc
Samuels, D. J. 2011. "The Modal Number of Citations to Political Science Articles Is Greater than Zero: Accounting for Citations in Articles and Books." *PS: Political Science and Politics*, 44(4): 783–792.
Sanchez, O., and A. Koukas. 2024, October 2. Departments Still Lack Unified ChatGPT Policy, Continue Adapting to Generative AI. *The Daily Princetonian*. https://tinyurl.com/5c5duytb
Saunders, L. 2011. *Information Literacy as a Student Learning Outcome*. Bloomsbury Publishing.
Selwyn, N. 2021. *Should Robots Replace Teachers? AI and the Future of Education*. Polity.
Slocum-Schaffer, S. A., and R. E. Bohrer. 2021. "Information Literacy for Everyone: Using Practical Strategies to Overcome 'Fear and Loathing' in the Undergraduate Research Methods Course." *Journal of Political Science Education*, 17(suppl 1): 363–379. https://doi.org/10.1080/15512169.2019.1694935
Sparks, J. R., I. R. Katz, and P. M. Beile. 2016. "Assessing Digital Information Literacy in Higher Education: A Review of Existing Frameworks and Assessments with Recommendations for Next-generation Assessment (ETS Research Report Series No. RR-16-32)." *ETS Research Report Series*, 2016(2): 1–33.
Spindel, J. and J. Ackerman. 2025. "Adapting to the AI Era: ChatGPT in the Classroom." *Journal of Political Science Education*, 21(2): 310–326. https://doi.org/10.1080/15512169.2024.2414017
Steele, M., B. McKibben, and J. Morinville. 2025. "Your Electric Bill Is Subsidizing Big Tech's AI Dreams | Opinion." *Pennlive*. September 23, 2025. https://tinyurl.com/yd2zy7z9

PROOFS

Strubell, E., A. Ganesh, and A. McCallum. 2019. "Energy and Policy Considerations for Deep Learning in NLP." October 5, 2023. http://arxiv.org/abs/1906.02243

Tong, S. T., A. DeTone, A. Frederick, and S. Odebiyi. 2025. "What Are We Telling Our Students about AI? An Exploratory Analysis of University Instructors' Generative AI Syllabi Policies." *Communication Education*, 74(3): 261–282. https://doi.org/10.1080/03634523.2025.2477479

Topal, R. and F. Shargh. 2023. "Teaching Students How to Find and Identify Reliable Online Sources: A Series of Exercises." *Journal of Political Science Education*, 19(3): 475–484.

Tuncel, O. 2025. "Teaching Literature Reviews in the Age of Generative Artificial Intelligence (AI)." *Political Science Educator*, 28: 2. https://tinyurl.com/3zk3t5j9.

Vallor, S. 2024. *The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking*. Oxford University Press.

Van Laar, E., Alexander J. A. M. van Deursen, Jan A. G. M. van Dijk, and Jos de Haan. 2017. "The Relation between 21st-Century Skills and Digital Skills: A Systematic Literature Review." *Computers in Human Behavior*, 72: 577–588.

Wallach, H. 2015. "Computational Social Science: Toward a Collaborative Future." In Alvarez R. Michael (Ed.), *Data science for politics, policy, and government* (pp. 307–316). Cambridge University Press. https://tinyurl.com/2z7yzch6

Walsh, J. D. 2025. "Everyone Is Cheating Their Way through College." *New York Magazine*, May 7. https://tinyurl.com/y2mr8kuu

Wargo, K., and B. Anderson. 2024. "Striking a Balance: Navigating the Ethical Dilemmas of AI in Higher Education." *EDUCAUSE Review*. https://tinyurl.com/24smb2jv

Wells, Q. 2025. "As AI Becomes Part of Everyday Life, It Brings a Hidden Climate Cost." *Yahoo News*. September 23, 2025. www.yahoo.com/news/articles/ai-becomes-part-everyday-life-130145069.html

Wiley, D. 2016. "Toward Renewable Assessments." *Improving Learning*, July 7. https://opencontent.org/blog/archives/4691

Williamson, B. 2016. Digital education governance: An introduction. *European Educational Research Journal*, 15(1): 3–13. https://doi.org/10.1177/1474904115616630

Wu, C. J., R. Raghavendra, U. Gupta, B. Acun, N. Ardalani, K. Maeng, and K. Hazelwood. 2022. "Sustainable AI: Environmental Implications, Challenges and Opportunities." *Proceedings of Machine Learning and Systems*, 4: 795–813.

Wu, N., and Patrick Y. Wu. 2024. "Surveying the Impact of Generative Artificial Intelligence on Political Science Education." *PS: Political Science and Politics*, 57(4): 602–609. https://doi.org/10.1017/S1049096524000167

Zuboff, S. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (First ed.). PublicAffairs.

PROOF

PROOFS

Index

18th Airborne Corps, 268
21st Century Town Meeting, 42
75th Innovation Command (US Army), 268

Aadhaar (India), 114–118
AAPOR. *See* American Association for Public Opinion Research (AAPOR)
AAPOR Task Force on Responsible AI Integration in Survey Research Report, 97
ABMs. *See* agent-based models (ABMs)
ABVV-Metaal, 39
academic integrity, 384
accountability, 110–111, 286, 287
  diagonal, 110
  frameworks vs. responsibility practices, 110–111
  horizontal, 110
  vertical, 110–111
Acemoglu, Daron, 27, 211, 218
active learning, 379
Ada Lovelace Institute, 129
Adams, Eric, 28
administrative burden, 159, 165, 180
Administrative Procedure Act of 1946, 118
Advanced Semiconductor Materials Lithography (ASML), 259
Afghanistan, 267
African American English, 155, 162
African American Vernacular English (AAVE). *See* African American English
afterlife technologies, 290
agent-based models (ABMs), 333
agentic AI, 4–9, 18, 23
AGI (Artificial General Intelligence). *See* Artificial General Intelligence (AGI)
agonistic democracy, 295
AI. *See* artificial intelligence (AI)
AI agents, 56, 87. *See also* agent-based models (ABMs); agentic AI
AI caregiving devices, 200
AI chatbots. *See* chatbots
AI detection tools, 382
AI ethics. *See* ethics
AI labor, 288. *See also* labor
AI literacy, 14, 381–384
AI Measurement Statement (AIMS), 11–17, 172–175
  core components, 172–174
  disclosure template, 172–175, 181–182
  four-question framework (table), 172
  illustrative examples, 174–176
  relationship to existing standards, 172–175, 177–179
AI penalty in deliberation, 294
AI representatives, 293–294
AI Safety Institute, 272
AI search, 55
AI slop, 54
AI Steve, 31
AI-administered surveys, 170
AI-assisted gerrymandering, 164
AI-assisted grading, 366–370
AI-assisted interviewing, 85–87
AI-assisted responding, 87
AI-free vs. AI-integrated assignments, 366, 377

PROOFS

Tucker_9781047774734_Index.indd 397 15-09-2026 07:09:33

AI-generated bots, 82. *See also* chatbots
AI-generated content, 57–58
AIMS. *See* AI Measurement Statement (AIMS)
Aldahoul, Nouar, 344
A-level grading algorithm (UK), 124
algorithm registers, 127–128
algorithmic bias
  in AI systems, 11–17
  in alignment (SFT, RLHF), 153
  in credit allocation, 198
  in democratic politics, 37
  in healthcare, 198–200
  in hiring and recruitment, 198–200
  ideological, 61
  in image recognition, 198, 199
  limits of technical fixes, 153
  machine bias, 347
  in political theory, 283, 300
  in pretraining data, 153
  in public opinion research, 91
  in racial politics, 153
  in research methods, 328–333
  six domains (Smith and Rustagi), 198
  structural and sociotechnical, 153
  in teaching and pedagogy, 384
algorithmic coloniality, 289
algorithmic curation, 55
algorithmic fairness, 116–121, 156, 283
  impossibility theorems, 156
  statistical parity conceptions of, 283
  structural turn in, 283–284
algorithmic fidelity, 91–92, 328
algorithmic governance, 137
algorithmic injustice, 13, 283–285
algorithmic management, 219–220
algorithmic transparency, 129
AlgorithmWatch, 126, 129
alignment. *See also* value alignment problem, 114, 136
All Eyes on Rafah image, 38
Alphabet, 222
Alqahtani, Naifa, 370
Altman, Sam, 6, 53
Amazon, 14, 222–224
Amazon Rekognition, 175
Amazon Web Services (AWS), 18
American Association for Public Opinion Research (AAPOR), 97
American Community Survey, 226
American National Election Study (ANES), 329
American Political Science Association (APSA), 13, 177, 179, 366
  Educate platform, 366
  membership survey on AI, 366–369
  Presidential Task Force on Rethinking Political Science Education (2024), 366–369, 370, 378
  Teaching and Learning Conference (TLC) survey, 366–369
AmericaSpeaks, 42
Amodei, Dario, 53
amplification of content, 297
amplification process (Ridgeway), 201
ANES. *See* American National Election Study (ANES)
Anno, Takahiro, 32
annotation and measurement, 313–316
  of images, audio, and video, 313
  validation of, 313–316
Anthropic
  in economic context, 224
  in information ecosystem, 53, 61
  in national security, 251, 259
  in public sector, 117
  as research tool, 313, 316
  in volume overview, 4–9
Anthropic Economic Index, 5
anti-democratic behavior, 229
anti-discrimination law. *See* civil rights law and algorithmic systems
antitrust policy, 235, 298
API. *See* application programming interface (API)
Apple Card, 200
application programming interface (API), 313–316
APSA. *See* American Political Science Association (APSA)
Arab Spring, 7, 81
Ardoin, Phillip J., 378
Argyle, Lisa P., 90–91, 322
Aristotelian motivation (LLM study), 343
Arizona State University, 366
arms control, 260–266
  dead zone, 270
Army Intelligence Data Platform, 267
Arrow, Kenneth J., 78
artificial discretion, 134
Artificial General Intelligence (AGI), 2, 222, 253
artificial intelligence (AI)
  asymmetric effects, 53–54
  definition of, 1–2
  as dependent variable, 195
  diffusion of capabilities, 247
  dual-use character of, 247, 249, 250, 254

PROOFS

effects on information ecosystem, 53–54
embodied, 3, 209, 216
exposure measures, 209–213
as general-purpose technology, 1, 2, 209, 216, 247, 249
historical analogs for, 1
as independent variable, 195
labor market effects, 209–213
as normal technology, 209, 216
political consequences of, 209–213
in political science research (overview), 313–314
public attitudes toward, 1
public opinion on, 209–213
regulation of, 1–2, 53
source vs. content effects, 53–54
Asian American communities, 160
asset specificity, 224
assignment design, 366–370
AI-free vs. AI-integrated options, 366, 377
disposable vs. renewable assignments, 366, 377
multi-modal submissions, 366, 378
oral exams and in-class reflections, 366
scaffolded assignments, 366, 378
AUC (Area Under the Curve), 314
audio analysis, 319
audit experiments (of LLMs), 314
audit studies, 323–324
auditability, 351–353
audits, algorithmic, 160
authenticity infrastructure, 64
authoritarian regimes, 61–69
Automatic Signature Verification, 36–37
Automating Inequality (Eubanks), 158
automation, 211, 247, 269, 276, 340, 341, 387
bias, 251
reshoring and, 211–212
robot exposure, 211, 212, 216
autonomous vehicles, 2–8, 285
autonomous weapons, 252–256, 285
Autor, David, 211–212
Avatar Democracy, 293
Axelrod, Robert, 333–335

Bagley, Nicholas, 132
Bai, Hui, 58–59, 323, 338
Bail, Christopher A., 338, 380
Barocas, Solon, 157
Barrie, Christopher, 66, 316, 321, 325, 338, 339
batch processing, 316
Becher, Michael, 217
belief relevance, 65
benchmarks and evaluations, 19
Bender, Emily M., 344, 366
Benjamin, Ruha, 157–162
Bergbower, Matthew L., 373
Berglund, Lukas, 343, 346, 349
Bertrand, Marianne, 345
Bhashini, 28
Bhatt, Alia, 38
bias. *See also* algorithmic bias
acquiescence, 347
automation, 251
gender, 198–202
in LLMs, 53–59, 78, 84, 313, 314, 324, 349
social desirability, 86–92, 169, 330
Biden, Joe, 58, 260–266
Big Fund (China Integrated Circuit Industry Investment Fund), 259
Big Tech, 6, 12
political economy of, 289
Binz, Marcel, 330, 332, 350
biographical data collection, 340
biometric identification, 63, 159
Birhane, Abeba, 284, 285
Bisbee, James, 82, 91, 327–331, 332
Black Americans, 162
black box (AI), 328, 329, 344
Black Twitter, 170
Bletchley Park AI Safety Summit, 272
Blumenthal, Richard, 53
Blumer, Herbert, 79
body-doubling, 379
Boelaert, Julien, 327, 331, 347
Boix, C., 226, 227, 231, 232
Bol, Damien, 320
Bolukbasi, Tolga, 328
border surveillance, 159
Bostrom, Nick, 387
bot networks, 55
Boussalis, Constantine, 195, 196
Brayne, Sarah, 159
Brazil, 109
breadwinner backlash, 12, 203
Breiman, Leo, 343
Briggs, Ryan, 321
Brighton Pavilion constituency, 31
Broockman, David E., 30
Brucks, Melanie, 331
Bryce, James, 79
Brynjolfsson, Erik, 213–215
Brynjolfsson AI exposure measure (SML), 213
Buchanan, Ben, 263, 269
bundled treatments, 313

PROOFS

Buolamwini, Joy, 155, 156
Bureau of Labor Statistics, 215
Burke, Edmund, 79
Buskirk, Trent D., 83, 85, 88

Cahill, Christine, 372, 384
California, 33
campaign finance, 14
campaign studies, 319
Canada, 30
Cantwell-Chavez, Daniel, 371, 376, 379
Canvas (LMS), 368
capital
  concentration of, 210–211
  labor share vs., 210–211
  specificity of, 210
  universal basic capital (UBC), 210
carbon chauvinism, 299
care work, 113
career preparation, 366, 370
causal inference, 153, 332
CBRN risk, 3
Cederman, Lars-Erik, 334
Center for Countering Digital Hate, 126
Center for Election Innovation and Research, 36
Centre for Responsible Union AI, 38
chat interfaces, 317
chatbots, 2–3, 55, 82
  citizens' private use of, 29–35
  in election administration, 29–35
  griefbots, 290–291
  in political campaigns, 29–35
  in social movements, 29–35
  voter preferences and, 82
ChatGPT
  CO_2 emissions, 367
  conflict mediation use, 367
  in democratic politics, 32–38
  in economic context, 216
  in gender politics, 199
  in information ecosystem, 53–56
  in national security, 253
  in political theory, 296
  in public sector, 109
  as research tool, 317–320
  in teaching, 366–370
  in volume overview, 2
ChatGPT 3.5 turbo. *See* GPT-3.5
cheating, 372–377
Cheating Lessons (Lang), 377
Cheo (chatbot), 39
Chile, 319
China, 5–10
  AI development strategy, 248–251
  semiconductor industry, 248–251
  state control over LLMs, 54–57
  US-China agreement on human control over nuclear weapons (2024), 254
  US-China AI competition, 254
Chinese Communist Party (CCP), 57
Chinese state media control, 96
Cho, Wendy K. Tam, 152
chokepoints (in AI supply chain), 254–258
Chouldechova, Alexandra, 156
Chuang, Yun-Shiuan, 336–338
Circular A-76 (Office of Management and Budget), 118
citizen deliberation, 26–27
citizen science, 129
citizens' assemblies, 27, 295
civic education, 367
civic engagement, 367, 387
civil discourse, 380
civil rights law and algorithmic systems, 160
civil society organizations, 37–41
classification tasks, 20
Claude (AI model), 4, 35, 61, 251, 259, 296
Claude Code, 4–9
Clinton, Hillary, 30
closed-weight models, 331
cloud computing, 222–224
Code for America, 132
code-switching, 163, 170
Codex, 4, 8
Coeckelbergh, Mark, 292, 300
Coglianese, Cary, 122–128
Cohen, Joshua, 295–299
collective bargaining, 210–214
collective self-determination, 294
colonialism, digital. *See* digital colonialism
Colorado Anti-Discrimination in AI Law, 18
Columbia University workshop (2024), 35
command and control, 249
COMPAS, 18, 111, 126, 168
compensation for AI harms, 285
complementary skills, 212, 216
CoMPosT framework, 96
computational map evaluation, 164
computational resources (inequality of access), 333
compute infrastructure, 13, 247–249
compute power, demand for, 211
computer vision, 174, 196
confederate (in experiments), 322–323

PROOFS

confidence-building measures, 275
Confucian philosophy, 285
confusion matrix, 320
consensus in deliberation, 295
consent, 290
conspiracy beliefs, 82
conspiracy theories, 58–59
construct fidelity, in experimental treatments (LLM-generated), 322
construct validity, 85, 154–158, 313–314
content moderation, 55, 64, 161, 296–297
contracting state, 117–123
conversational agents, 169, 325
Converse, Philip E., 80
copyright, 374
corporate profits, 212
Costello, Thomas H., 58–66
Couldry, Nick, 375
Cowan, Mike, 372
Crawford, Kate, 288–289, 375
criminal justice, 159, 285–287
Critical and Emerging Technology Working Group (Quad), 274
critical mass theory, 197
critical race theory, 284
critical thinking, 365–368, 372, 376
cross-model triangulation, 94
cross-racial coalition, 162
crowdsourced adaptive surveys (CSAS), 65
crowdworkers/crowd work, 324
cultural imperialism, 289
cultural smothering, 285
curriculum design, 374
customer support, 217, 232
Customs and Border Protection (CBP), 60
cyber operations, 249–254
cyberspace, 248–252

Dahl, Robert, 41–42
Daily Princetonian, The, 370
DALLE, 38
Danaher, John, 286, 292
data access, 5, 177
Data Access and Research Transparency, 177
data centers, 6, 219, 224, 248–251, 288
data colonialism, 289
data creation, 195–196
data donation, 9
data leakage, 333, 339
data poisoning, 55
data privacy, 28, 374–375
data voids, 62
data workers, 288
Databricks, 224
datafication, 375
Dawson, Michael, 162
De Maio, Jennifer, 373, 381
deadbots, 290
Dean.Bot, 32
de-biasing methods, 321
decision-support systems (military), 249
decolonial studies, 289
Deep Thought (fictional computer), 78, 99
deepfakes, 10, 28, 53, 54. *See also* AI-generated content
  of deceased people, 290
  deepfake pornography, 201
  gender-based targeting, 201–202
DeepSeek, 19, 256, 257, 313
Defense Production Act, 61
deliberation, 10–13
Deloitte, 117, 119
Delphi, 32
democracy, 10, 282–286
democratic backsliding, 229, 233
Democratic Party (US), 210
democratic publics, 295
democratic values, 291–293
democratization of AI, 292
Department of Government Efficiency (DOGE), 126
departmental policies (AI), 365
DeSantis, Ron, 263
deterrence, 252–257
developing economies, 210–215
Dewey, John, 291, 295, 297
Diamond, Larry, 6
difference principle, 284
differential privacy (Census), 171
digital authoritarianism, 53
digital colonialism, 289–290
Digital Dialogues (Libya), 44–47
digital divide, 377
digital language divide, 289
digital literacy, 366–370
Digital Markets Act (European Union), 133
digital public sphere, 292–298
digital self-determination, right to, 288
Digital Services Act (EU), 3
digital twins, 327, 329
Dilling, Matthew, 381
disclosure effects, in experimental treatments (LLM-generated), 322
discretion, 115
disinformation, 53–61
disposable vs. renewable assignments, 366, 377

PROOFS

distributive justice, 13, 283, 284
diversified quality production, 218
Doellgast, Virginia, 218–220
Dole, 15
Donahoe, Emily Pitts, 376–379
door-to-door canvassing, 29
drone warfare
  first-person-view (FPV) drones, 251
  non-state actor use, 251
  in Russia-Ukraine war, 251–255
drones, 3. *See also* drone warfare
Duch, Raymond, 92–94

East Asia, 231
echo chambers, 54
economies of scale and scope, 223
education and training, 210, 212, 234. *See also* training data
effective accelerationism, 228
Egami, Naoki, 169, 321, 329
Eggert, Linda, 287
elections, 10
  2016 US presidential, 54
  2020 US presidential, 36
  2024 Indian, 28
  2024 US, 54
  2024 Venezuelan, 39
  administration, 10, 26–27, 29–36. *See also* electoral administration
  interference, 53
electoral administration
  polling place allocation, 153
  signature matching, 153–158
  voter roll purges, 153
electoral dealignment, 226
electoral realignment, 210, 211, 226
Electric Twin, 28
ElevenLabs, 28
Elsa (FDA AI tool), 116
embedded systems, 220
Endacott, Steve, 31
Enron emails, 375
environmental costs and impact of AI, 285, 373
environmental justice, 284–288
epistemic actorhood, 298
epistemic collapse, 79, 94
epistemic injustice, 284
epistemic justice, 13
Epstein, Joshua M., 333–335
equal basic liberties, 284
equality, 282
equity in grading, 366–370
escalation, 248–251
essay detection accuracy, 369
Estonia, 115
ethics, 322, 325
  in AI use, 366–370, 373
  non-Western traditions, 285
  Ubuntu, 285
Eubanks, Virginia, 158, 375
European Commission, 221
European Union (EU), 3, 33, 126, 157, 180. *See also specific countries*
  AI Act, 3, 33, 126, 157
Evsyukova, Yulia, 322, 324
Executive Order on AI policy (December 11, 2025), 19
existential risk from AI, 290
experimental treatments (LLM-generated), 322
explanatory motivation (LLM study), 343
export controls, 13, 249–252, 260
external validity, 313
extreme ultraviolet (EUV) lithography, 259
extremist networks, 62
Exxon, 15

F1 score, 331
Facebook, 7, 30, 55, 126, 296
  messenger, 30
facial recognition technology, 63, 130
  Amazon Rekognition example, 175
  Gender Shades study, 155
faculty development, 370, 372, 384
faculty surveys, 368
fair equality of opportunity, 284
fairness, algorithmic. *See* algorithmic fairness
Farrell, Henry, 195, 198, 254, 256, 295
feedback loops, algorithmic, 155
Felten, Edward, 213
Felten AI exposure measure (AIOE), 213
feminist theory, 283
few-shot prompting, 319
fine-tuning, 93, 318, 320
first-generation students, 380–382
Fishkin, James S., 79
FLOPs thresholds, 271
Florida, 36
Foos, Florian, 29, 34
foreign influence campaigns, 7
foundation models, 122–127, 210–214
Founding Fathers (US Constitution), 18
Fourcade, Marion, 195, 198
framing effects, 89–92
France, 320

PROOFS

Tucker_9781047774734_Index.indd 402 15-09-2026 07:09:33

Fraser, Nancy, 284, 295
Freedom of Information Act, 36, 128
Freedom of Information requests, 36
Freelon, Deen, 170
Frey, Carl Benedikt, 214, 221
frontier AI models, 12
Frontier AI Safety Forum, 272
Frontier Model Forum, 272
Fujitsu, 135
fundraising emails, 28
future generations. *See* intergenerational justice
Future Party (Japan), 32

GABMs. *See* generative agent-based models (GABMs)
Gabriel, Iason, 283–288
Gallup, George, 80
Gaza, 38
  war, 264
GDPR. *See* General Data Protection Regulation (GDPR)
Gebru, Timnit, 155, 156
Gemini (Google), 26–31
gender. *See also* gender and politics, study of
  AI and, 12, 155, 195, 202–204
  bias, 202–204
  ChatGPT and, 198
  deepfakes and, 201–202
  gap, 197
  gender-based violence, 12
  gendered data gaps, 202
  gendered division of labor, 195
  LLMs and, 198–202
  as primary frame, 198
  sex vs., 197
gender and politics, study of
  AI as category of analysis, 195–200
  AI as method of data analysis, 195
  AI as tool of data creation, 195–200
Gender Shades study, 155
General Data Protection Regulation (GDPR), 34
general-purpose technology, 247, 272
generative agent-based models (GABMs), 333–337
  memory modules in, 333–337
  orchestration and evaluation, 333
generative AI, 2. *See* large language models (LLMs)
  in democratic politics, 27
  in economic effects, 213, 221
  gender bias in, 202–204
  in information ecosystem, 53–59
  public concern about, 202
  in public sector, 110–111, 112
  in teaching and pedagogy, 366, 367
GeoSpy, 63
Germany, 216, 221, 264, 319
  German leadership debates, 196
gerrymandering, 164
Get-Out-The-Vote (GOTV) campaigns, 29–34
Gillum, Rachel, 375
Gilman, Nils, 374
GitHub, 56
GitHub Copilot, 217
Global Investigative Journalism Network, 40
global justice, 288
Global North/Global South, 285
Global Partnership on Artificial Intelligence (GPAI), 128
Global South, 5, 210–215, 371
global value chains, 219, 231
globalization, 210–215, 225
gold standard, 313, 321, 352
Goldfarb, Avi, 268
Goldstein, Josh A., 322, 324, 349
González-Rostani, V., 226
Google, 10, 14, 55, 223–224, 316, 318
Google DeepMind, 10, 26
GOTV. *See* Get-Out-The-Vote (GOTV) campaigns
GovAI Coalition, 133
Government of National Unity (Libya), 45
government use of AI, 294
  immigration enforcement, 157
  policing and surveillance, 155
  public administration and eligibility, 157
GPT-2, 325
GPT-3, 65, 258, 319
GPT-3.5, 319
GPT-4, 59, 319
GPT-4.5, 59
GPT-4o, 59, 65, 176, 325
GPT-5, 316
GPUs, 223
grading, 124, 366–370
griefbots, 290
Griffin, James, 298–299
Grimmer, Justin, 328
Grok, 58, 296
ground truth, 92–95
group consciousness, 161, 162
Guess, Andrew M., 54, 64, 383

PROOFS

Guevara, Freddy, 38
GUIDE-LLM, 17, 177
  checklist, 177
Gunkel, David, 300
Guterres, António, 53

Habermas, Jürgen, 295, 297
Habermas Machine, 10, 26–31, 295
Hackenburg, Kobi, 30, 58–59, 322, 325
Hadid, Bella, 38
Hall, Andrew, 2–4
hallucination, 20, 26, 68, 344, 349
Han, Sungmoon J., 282–285
Hanna, Alex, 366
harassment, online, 63
HAROP (Israeli loitering munition), 252
Harris, Kamala, 30
Harvard Law School, 366
Harvard Youth Poll, 221
hate speech, 296, 320
  detection, 320
Hawthorne effect, 330
Health Resources Action, 40
Herbst, Susan, 80
Herd, Pamela, 159
hermeneutical injustice, 284
Hersh, Eitan, 162–163
Heseltine, Michael, 319, 349
Hewitt, Luke, 325, 327, 332
high-resource languages, 56
Himmelreich, Johannes, 292
Hindi language, 56
Hispanic communities. *See* Hispanic populations; Latino communities
Hispanic populations, 171
Horizon (UK Post Office), 110
Horowitz, Michael C., 248–250
housing and algorithmic systems, 158
Hovy, Dirk, 155
Howard, Jeffrey, 296–297
H.R. 1 (One Big Beautiful Bill Act), 119
human benchmark, 352
human coders, 351–352
human rights, 53
  fourth generation (digital age), 298–301
  grounds of, 298–301
  inflation worry, 298
Huq, Aziz, 296
hyperscalers, 262

IBM, 11, 35, 119–124
IBM Watson, 35
ICE. *See* Immigration and Customs Enforcement (ICE)
Idaho, 35
image analysis, 314
image generation, 201–202
image recognition systems, 202
Imai, Kosuke, 66
immigrant communities, 164
immigration, 157, 221
  enforcement, 157
Immigration and Customs Enforcement (ICE), 60, 62
implicit racial inference, 166–172
imputation, 82–87
incidents-at-sea arrangements, 275
incompletely theorized agreements, 115, 137
India, 27, 57, 114–118, 265
Indiana, 6, 124
industrial robots, 232
industries and sectors
  agriculture, 215
  arts and entertainment, 213
  construction, 224
  educational services, 212
  finance and insurance, 215
  information sector, 215
  manufacturing, 211
  professional, scientific, and technical services, 215
  public administration, 210
  retail and wholesale trade, 215
  transportation, 212
influence operations, 61–63
information and power asymmetry, private vendors and, 109–112
information environments, 152–157
information superiority, 267
informed self-advocacy, 297
inoculation (against misinformation), 59
input batching, 318
Input-Process-Output (IPO) model, 292
Instagram, 38, 55
institutional racism, 287
Instrumentalist motivation (LLM study), 343
intelligence analysis with AI, 247
intelligence, surveillance, and reconnaissance (ISR), 250
intercoder reliability, 20, 315
intergenerational conflict, 227
intergenerational justice, 290–291
International Atomic Energy Agency (IAEA), 272

PROOFS

International Federation of Robotics (IFR), 231
international humanitarian law, 252, 264, 274
International Labour Organization (ILO), 217
International Political Science Association (IPSA), 388
International Studies Association (ISA), 369
International Telecommunications Union, 273
intersectional analysis, 181
intersectionality, 284
interviewer effects, 86
interviewing, AI-assited, 85–87
Iran, 251, 259
IRB considerations, 315
Israel, 38, 252, 264
Italy, 221, 319
item nonresponse, 80, 89

jagged intelligence, 68
Japan, 32
Jervis, Robert, 253, 269
job displacement, 3–8, 221–222
job polarization, 209–210, 211
Johnson, James, 254
justice. *See also* algorithmic injustice
  criminal, 159, 285–287
  distributive, 13, 282, 283
  environmental, 282–283
  epistemic, 13
  global, 282–283
  hermeneutical, 284
  intergenerational, 282–283
  rectificatory, 282–283
  theories of, 282–283

Kahl, Colin, 255
Kalla, Joshua L., 30
Kapoor, Sayash, 212, 216, 217, 233
Kappa scores, 315
Karpathy, Andrej, 68
Kasirzadeh, Atoosa, 284, 298
Keener, R., 372
Kehlenbach, E. Stefan, 371, 375
Kettl, Donald, 118–121
Key, V. O., 79
Khan, Imran, 31
Kim, Junsol, 330–334
KKR, 374
Kleinberg, Jon, 156, 159
knowledge production, 154
Kozlowski, Austin C., 327, 328, 331
Kremlin, 62

labor, 199. *See also* labor market; labor unions
  displacement, 3–8, 221
  exploitation, 282
  force participation, 195
  share of national income, 209, 210–211
labor market
  augmentation by AI, 209–210
  disruption by AI, 344
  effects, 209–213
  polarization of, 209–210, 211
  substitution by AI, 209–210
labor unions, 211, 217
  AI strengthening, 210–215
  AI use by, 39
  AI weakening, 210–215
Landemore, Hèléne, 294
Landgrave, Michelangelo, 323–324
Lang, James, 365
language models. *See* large language models (LLMs)
large language models (LLMs), 13, 17
  for annotation and measurement, 313
  bias in, 314, 324, 349
  in citizen deliberation, 30–36
  for coding open-ended responses, 78
  in democratic processes, 293–298
  in economic effects, 222–225
  in election administration, 30–36
  as gatekeepers, 53–59
  gender and, 198–202
  in gender politics, 198
  for generating survey items, 78–85
  for generative agent-based modeling, 313
  in geopolitics and national security, 258–263
  ideological bias of, 313
  for imputation, 78–85
  in information ecosystem, 53, 55, 56
  low-resource languages, 53–59
  as next-token prediction engines, 313
  as objects of study, 313
  open-source vs proprietary, 78–85
  outperforming humans, 313
  personalization in, 53, 55, 313
  persuasiveness of, 30–36
  political bias in, 53–59, 78, 84
  in political campaigns, 30–36
  post-training for persuasion, 53–59
  pretraining and racial bias, 154–155
  in public opinion research, 78
  in public sector, 112–115, 126, 129

PROOFS

large language models (LLMs) (cont.)
in racial politics, 154
as research agents for data collection, 313
as research infrastructure, 154–155
as research tools, 53–59, 313
as salary negotiation, 198–202
for silicon sampling, 313
as survey respondents, 78–85
in teaching and pedagogy, 368
for treatment and stimuli generation, 313
Larooij, Maik, 338
Lasswell, Harold D., 315
latent opinion, 80–84
Latin America, 231
Latino communities, 161, 170
Lavender (Israeli targeting system), 264
LAWS. *See* lethal autonomous weapons systems (LAWS)
LAWS Group of Governmental Experts (GGE), 251
Lazar, Seth, 282–286
learning management systems (LMS), 368
Lechterman, Theodore, 286, 293
Lee, Taeku, 1
legal responsibility for AI harms, 286
legitimacy, 43–47
Lerman, Amy, 159
lethal autonomous weapons systems (LAWS), 273–274
LGBTQ+ individuals, 201
liberal arts education, 367
Libya, 44–47, 266
Lin, Hause, 26
Lindsay, Jon, 268
Line of Actual Control, 265
Linked Fate, 162, 181
LinkedIn, 324
Lipa, Dua, 38
Lippmann, Walter, 79
listening campaigns, 30
literacy tests, 163
literature review tools, 172
literature reviews (teaching of), 373
Llama (LLM), 54
Llama 2, 35
LLM bias. *See* large language models (LLMs)
LLM-generated content. *See* AI-generated content
LLMs. *See* large language models (LLMs)
LM Studio, 317
lobbying, 210–214
local election officials (LEOs), 34–35
lock-in, private vendors and, 109–112
loitering munitions, 252, 274
Lonergan, Erica, 269
Los Angeles Police Department (LAPD), 63
Lovett, Frank, 294
low-resource languages, 56
Lyman, Alex, 331

machine learning, 2, 8, 196–197, 220
classifiers, 296
double, 66, 168
machine learning models. *See* machine learning
machine translation, 158, 168, 174
Macron, Emmanuel, 30
mail-in ballots, 36
malicious uplift, 266
manipulation, 285
Manuali, Lorenzo, 293
Marfleet, B. Gregory, 383
Margetts, Helen, 322
Maricopa County, Arizona, 36
Marx, Karl, 14
Mass Online Deliberation, 294
Massachusetts, 30
matplotlib, 56
Maven Smart System, 251, 268
mBERT, 175
McKinney, Claire, 295
meaningful human control, 252
measurement error, 79, 166–172
median voter, 229
Medicaid, 119
Mellon, Jonathan, 83, 318–320
Merkel, Angela, 196
Meta (company), 7, 29, 313
Mettler, Suzanne, 160, 165
Mexico, 27
Miami-Dade Sheriff's Office (MDSO), 63
Michels, Steven, 377–381
Michigan, 11, 36, 123
Michigan Integrated Data Automated System (MiDAS), 123–127
Microsoft, 14, 216–220
microtargeting, 62, 162, 201
MiDAS. *See* Michigan Integrated Data Automated System (MiDAS)
MiDAS unemployment system, 11
middle-income trap, 233
Midjourney, 38
Mikhaylovskaya, Anastasia, 295
military organizations, 203

PROOFS

Miller, Chris, 256, 259
mini-publics, 294
Minneapolis, 60
minority political participation, 162–167
misinformation, 3, 34–36, 54, 58, 161, 349
Mistral, 35
MIT report on AI adoption, 366
Mitchell, Margaret, 166
model cards, 177
model collapse, 93. *See also* epistemic collapse
model deprecation, 313
model drift, 66
Model United Nations, 372
model versioning, 313
Modi, Narendra, 28
Moltbook, 56
Monsanto, 15
moral rights, 299–300
moral status, 299
mortgage models, 158
motivated reasoning, 64–65
Motoki, Fabio, 344, 349
Mouffe, Chantal, 295
Moulton, Jeremy F. G., 386
Moynihan, Donald, 159
Muldoon, James, 288
Mullainathan, Sendhil, 345
multidimensional treatments, 322–324
multilingual research, 167
multi-modal submissions, 366, 378
Munger, Kevin, 7, 60, 81–85, 349, 383
Mythos model, 5

Narayanan, Arvind, 212, 216, 217, 233
NASUWT, 39
National Institute of Standards and Technology (NIST), 110, 153
National Science Foundation (US), 110
national security, 2–8
National Security Commission on AI, 255
NATO, 263
natural language processing (NLP), 167, 197, 346
natural monopoly, 223
Nellis, Mike, 27
Netherlands, 216
Nevada, 35
New Deal, 118
New Public Management, 113
New York City, 28
New York State Department of Financial Services, 200
New York State Senate Bill S9678, 28
Newman, Abraham, 254, 256
news media
  AI impact on, 54
  in algorithmic age, 298
NIST. *See* National Institute of Standards and Technology (NIST)
Noble, Safiya, 154, 167, 172
non-identity problem, 291
nonprobability samples, 83
nonresponse. *See* item nonresponse
non-state actors, 270
normal technology view of AI, 16
North American Industry Classification System (NAICS), 215
nuclear weapons, 13
NVIDIA, 14, 223, 261

O*NET, 215
occupational segregation, 202–204
Occupy Wall Street, 7
October 2022 export controls, 260
OECD. *See* Organisation for Economic Co-operation and Development (OECD)
offense-defense balance, 249
Offer-Westort, Molly, 30
Office of Management and Budget, 118
offshoring, 222–225
oil-spill model of politicization, 81
older workers, 212, 232
Ollama, 317
Omi, Michael, 155
online harassment, 201
online public shaming, 297
online surveys, 87, 93
opacity
  of AI models, 297
  algorithmic, 160
open source AI, 54
OpenAI
  in democratic politics, 32–37
  in economic context, 224
  in information ecosystem, 53–59
  in national security, 258
  in racial politics, 177
  as research tool, 313, 340, 349
  in teaching, 374
  in volume overview, 6, 14
open-ended survey responses, 83–88, 320
openrouter, 316
open-source AI models, 248
Open-Source Intelligence (OSINT), 63

PROOFS

open-source tools, 41
open-weight models, 316–320, 352
Operation Spider's Web, 248
opinion dynamics models, 332
oral assessments, 368
oral exams and in-class reflections, 378
Organisation for Economic Co-operation and Development (OECD), 111, 116, 121, 217, 219
Ornstein, Joseph T., 319
overreliance on AI, 45
Owen, E., 226

Packback, 382
Pahlka, Jennifer, 130–132
Pakistan, 31
Pakistan Tehreek-e-Insaf (PTI), 26
Palmer, Alexis, 96, 316, 321
pan-ethnic identity, 162
parasocial relationships, 59
Parfit, Derek, 291
Park, Joon Sung, 330–334
participation washing, 284
participatory approaches to AI, 284
Pascal, Pedro, 38
PDF parsing, 341
pedagogy, 10–14
  algorithmic bias in, 384
  generative AI in, 365–367
  LLMs in, 368, 369
peer review, 23
Persian Gulf War (1991), 267
Persily, Nathaniel, 5
persona prompts, 329
personally identifiable information, 318
personas, 79, 85, 91
persuasion, 53, 54
  AI-driven, 325, 333, 335, 337
  personalized, 53, 54
  role of facts and evidence, 53–55
p-hacking. *See* prompt-hacking (p-hacking)
Phillips, Dean, 32
physicians, AI-generated images of, 202
Picón, Isabella, 40
Pindyck, Shira, 203
placebo control stories, 324
plagiarism, 367–371
platform APIs, 170
platform workers, 222
pluralistic ignorance, 81
Pol.is, 43–45
Poland, 30
polarization. *See* job polarization; political polarization
political campaigns, 5, 26–32, 152, 162
  chatbots in, 29–35
  listening and dialogue, 26–27, 30–32
  LLMs in, 30–36
  message development, 28
  microtargeting and propensity modeling, 162–163
  mobilization, 26–27, 29
  personal data regulation and, 33–34
  persuasion, 26–27, 29–30
  voter suppression risks, 152
Political Declaration on Responsible Military Use of AI and Autonomy, 274
political polarization, 3, 7, 32–38, 54, 201, 210–215, 229, 297, 337–338
political representation, 171, 293
political theory, 13, 282
  algorithmic bias and, 283, 301
  ChatGPT and, 295
  as normative enterprise, 282
poll taxes, 163
populism, 233
Porter, Ethan, 322–326
Porto Alegre water meter bill, 109
Post Office (UK), 125
poststratification, 81
post-training alignment, 90, 91
power, 195
Pravda network, 62
predictive AI, 367, 375
predictive policing, 159, 287
preference collapse, 156
Presidential Task Force on AI, Politics, and Political Science, 1
pretesting, 82–85
primary frame (Ridgeway), 198
Princeton University, 370
privacy, 28, 171, 290, 374–375
private vendors, 110
probability samples, 83
procurement, 111–114
productivity, 209–212
project-based learning, 378–379
prompt caching, 316
prompt design, 85, 88
prompt engineering, 338
prompt-hacking (p-hacking), 347
prompts
  design and sensitivity of, 317, 324
  pre-registration of, 317

PROOFS

system, 317, 324
user, 317
propaganda, 57, 59, 61, 324
proprietary AI models, 94, 338
ProPublica, 129
protest, 7
provenance and logging, 314
proximity voting models, 65
public administration, 157
Public and Commercial Services Union (PCS), 39
public opinion, 7, 10, 21, 78, 97
on AI, 209–213
AI effects on, 78–79
concept of, 78–79
definitions of, 78
democratic legitimacy and, 78–79
public records requests, 36
public service values, 110
public sphere. *See* digital public sphere
punishment, 285–288

Quad (the), 274
qualtrics, 93
question wording, 80–86
questionnaire design, 84
Quiller, 28

race, 11, 198, 375
critical race theory, 283
LLMs and, 154–155
pretraining and racial bias, 154–155
in racial politics, 154
as research infrastructure, 154–155
racial formation theory, 152
racial group consciousness. *See* group consciousness
racial inequality, 375
racial proxies, 166
radicalization, 62
Rafah, 38
RAG. *See* retrieval-augmented generation (RAG)
Rainbow CH-4 (drone), 266
random forests, 343
ranking algorithms, 338
Raven Sentry (program), 267
Rawls, John, 283
Raymond, Chad, 365–369
Reagan administration, 118, 253
reasoning parameter, 314
rectificatory justice, 285
redistribution, 221–225, 229
redistricting, 163, 164
redlining, 158
regulation, 1–2, 53
European Union AI Act, 3, 33, 126, 157
General Data Protection Regulation (GDPR), 33
regulatory capture, 12, 230–233
reinforcement learning from human feedback (RLHF), 155
Reiter, Dan, 254
relational egalitarianism, 284
Remesh AI, 44
replicability, 16–17, 66–68
replicate.com, 316
replication
of AI-integrated research, 95–98
in LLM-based research, 314–317
transparency and documentation, 314–317
reproducibility, 13, 17, 94–96, 314–315
Republican Party (US), 209
reshoring, 231–235
responding, AI-assisted, 86
responsibility for AI-inflicted harm, 286
responsibility gap, 13, 286–287
Responsible AI in the Military Domain (REAIM), 274
Restrepo, Pascual, 211–219
résumé-screening algorithms, 155
Rethinking Political Science Education (APSA, 2024), 366–367
retribution, 287
retrieval-augmented generation (RAG), 62, 68, 113–117
revenge porn, 201
reward hacking, 114
Ridgeway, Cecilia, 198, 201
right to a human decision, 296
rights, 282–286
civil rights law and algorithmic systems, 160
human, 53, 298–301
moral, 299–300
robot, 298–301
voting, 164
Ringmar, Erik, 372
risk assessment
algorithmic, 287
in criminal justice, 159
tools, 111
Risse, Mathias, 282–286
RLHF. *See* reinforcement learning from human feedback (RLHF)

PROOFS

robot rights, 298–301
robot tax, 229
robotics, 3, 217, 219, 233
robustness checks, 336–339
Rohingya refugees, 88
Roman Catholic Church, 285
Routine Task Intensity (RTI) index, 209
routine vs. non-routine tasks, 209
Rozado, David, 58, 346, 349
rubrics, 382–384
Russia, 59–61, 265
Rustagi, Ishita, 198–200

Sagan, Carl, 299
Salvi, Francesco, 58, 344, 349
Sanders, Bernie, 263
Sanders, Nathan E., 27, 32, 40
Santurkar, Shibani, 82–86, 327–333
satisficing, 89–92
Saudi Arabia, 258, 266
scaffolded assignments, 366, 378
scale creation, 325
scaling logic (in AI development), 254–260
Scarlet Dragon (program), 268
Schelling, Thomas C., 334
Schelling segregation model, 334
Schneier, Bruce, 27, 32
scientific polling, 80
search engines, AI-powered, 282
sector-specific exposure measures, 215
security dilemma, 13, 248–252
Selbst, Andrew, 157, 160
semiconductors, 13, 249
Sen, Maya, 152
sensemaking, 42–47
sentiment analysis, 167, 197, 316
sex vs. gender, 197
Shambaugh, Scott, 56
Shanahan, Murray, 344, 350
silicon respondents, 79, 83
  algorithmic fidelity, 79–83
  commercial providers, 79
  demand effects, 79
  evaluation criteria, 79–83
  hybrid designs, 79–83
silicon sampling, 13, 313–316, 327–333
Simulmatics Corporation, 78
Singapore, 264
skill-biased technological change, 218
Skynet-Terminator scenario, 3
smart-buyer capacity, 120
Smith, Adam, 14
Smith, Genevieve, 198–200
social choice theory, 80
social desirability bias, 86, 169, 330
social media, 54–59, 295–296. *See also specific platforms*
  data access, 2–5, 314
  as liberation technology, 3
  manipulation, 54–59
  as paradigm for AI analysis, 2–5
  political research and, 154
  public opinion and, 78–82
social movements, 10, 26–27
Social Security Administration (US), 118
softmax function, 317
software development, 216–220
Sora, 38
South Africa, 27
South Korea, 53, 264
sovereign wealth fund, 234
Spanish language, 161
Spindel, Jennifer, 372, 376, 381
spiral model, 254
Spirling, Arthur, 96, 316
St. George's Hospital Medical School, 200
Stack Exchange, 56
Stack Overflow, 56
Standard Occupational Classification (SOC), 209
standards-setting, 273
Stanford University, 22, 124
state capacity, 110, 112
statistical discrimination, 287
statistical validity, 313
steerability of LLMs, 329
Stegmueller, Daniel, 217
STEM workforce, 212
stock wealth, US households, 230
Strategic Defense Initiative, 253
street-level bureaucrats, 119, 123, 134, 135
structural injustice, 284
structural power of capital, 230
student learning outcomes (SLOs), 383–384
student surveys (on AI), 365
Sunstein, Cass, 115, 137, 297
supervised fine-tuning (SFT), 155
surveillance, 26, 37, 63, 372
  border, 159
  carceral, 162, 169
  government use of AI and, 155
  intelligence, surveillance, and reconnaissance (ISR), 250
  state, 153

PROOFS

surveys
AI-administered, 169
crowdsourced adaptive surveys (CSAS), 65
experiments, 210, 323
faculty, 368
online, 87, 93
open-ended survey responses, 83–88, 320
research, 78, 97
student (on AI), 365
SUTVA (Stable Unit Treatment Value Assumption), 348
syllabi policies on AI, 376
synthetic audiences, 28
synthetic data, 313–316
synthetic media. *See* AI-generated content
synthetic respondents. *See* silicon respondents
system cards, 177

tailored experiments, 64–66
Taiwan, 10, 26–31
Taiwan Semiconductor Manufacturing Company (TSMC), 258–259
Talk to the City, 43, 45
Tan, Jonas, 209, 210
Tang, Audrey, 43
tariffs, 229
task model of production, 211
task-based framework, 26
Taxonomy of AI use by governments, 110
Taylor, Christopher, 286
teaching. *See* pedagogy
Teaching and Learning Conference (TLC), 368–369
techno-racism, 157, 159
temperature (model parameter), 17, 317–321, 348
temporal validity, 7, 60, 96, 313
tenant screening, 158
testimonial injustice, 284
text analysis, 88
text classification, 314
text confounding, 66–67
text-as-data, 8, 196
text-to-image generation, 202
Thelen, Kathleen, 222, 224, 231
Third-Party Doctrine, 374
TikTok, 38, 55
Times, The (London), 28
tokens, 316, 317, 318
Tokyo, 32
Toma El Control (TEC), 26
tool-augmented data collection, 314
topic modeling, 197, 346
tort liability for AI, 285
Total Survey Error framework, 83–91
Toubia, Olivier, 327, 330
town meetings, 41
toxicity detection, 167
trade protectionism, 226
trade secrets doctrine, 124
trade unions. *See* labor unions
Trager, Robert, 273
training data, 89, 90, 315, 318, 319, 323
translation, AI-enabled, 28
transparency, 110–113, 121, 288–293
in AI-assisted research, 88–95
in LLM research, 314–317
Treaty on the Prohibition of Nuclear Weapons (TPNW), 274
trolling, 62
Trump, Donald, 30, 58, 59, 260, 261
administration, 61, 119
trust in institutions, 160
trustworthy AI, 285
TSMC. *See* Taiwan Semiconductor Manufacturing Company (TSMC)
Tucker, Joshua A., 5, 8
Tuncel, Ozlem, 373
Turner, Charles C., 372
Twitter (X), 14, 81, 170, 295, 338
Tyler, Tom, 160

Uber, 43
UBI. *See* universal basic income (UBI)
Ukraine, 2, 265
unit nonresponse, 82
United Arab Emirates, 258, 266
United Kingdom, 27, 31, 38, 44, 128, 216
A-level grading algorithm, 128
Horizon Post Office case, 128
United Nations, 44, 53
UN Convention on Certain Conventional Weapons (CCW), 273
UN High-Level Advisory Body on AI, 272
United Nations Security Council, 53
United States Digital Service, 132
United States, 2–7, 15, 27–33, 34. *See also specific states and governmental entities*
AI exposure by income decile, 216
AI procurement and regulation, 53
AI strategy and policy, 248
data center infrastructure, 248–251
federal AI use cases, 111, 123, 132
labor share, 216–222

PROOFS

United States (cont.)
law enforcement use of AI, 62
partisan vote and AI exposure, 216, 221
universal basic capital (UBC), 218
universal basic income (UBI), 234
University of California, Riverside, 375
US Census, 171
US Department of Defense, 251
US Department of War, 251
US Food and Drug Administration (FDA), 116
US Government Accountability Office (GAO), 111–113, 125
US Senate, 319
Committee on Finance, 202
US State Department, 36
US-China agreement on human control over nuclear weapons (2024), 254
US-China AI competition, 254

Vafa, Keyon, 329
validation, 313–314
value alignment, 282, 387
problem, 387
Vaynman, Jane, 250, 270
Velez, Yamil, 64–68, 322–325
vendor lock-in, 119
Venezuela, 38, 39
vertical integration, 224
video analysis, 314
Vietnamese language, 161
Virginia, 36
Volpe, Tristan, 250, 270
voluntary commitments (AI), 272, 276
voter file, 162
voter fraud claims, 28
voter suppression, 162–164
voting behavior, 7
Voting Rights Act, 164
Vredenburgh, Kate, 297
vTaiwan, 26–31

wages, 210–214
bargaining, 210–215
inequality, 226
premium, 216, 224
Wagner, Gerit, 172
Waight, 57, 61
Wake County, North Carolina, 36
Walsh, John D., 382
warfare, 2
warrior identity, 203
Wasow, Omar, 168
water consumption (AI infrastructure), 375
Weaver, Vesla, 159, 169
web panels, 87
Webb, Michael, 213
Webb AI exposure measure, 213
Wei, Jason, 318
weighting, 92
welfare and social services, 158
welfare states, 235
Western Europe, 219, 226. *See also* European Union (EU); *specific countries*
Western hegemony, 289
Westwood, Sean J., 82–86, 327, 349
WhatsApp, 39, 55
Whiteness of AI, 289
whole-society simulation, 337
Wiley, David, 377
Winant, Howard, 158, 166
Wing Loong II (drone), 266
Winning the AI Race (America's AI Action Plan), 255
woke AI, 18
women, in politics, 195, 197
workslop, 374
World Bank, 232
World Values Survey, 329
writing assignments, 366
Wu, Patrick Y., 319

X (Twitter). *See* Twitter (X)
xAI, 224
Xi Jinping, 255

Yemen, 266
Yin, Yidan, 315–318
Yoon Suk Yeol, 53
Young, Iris Marion, 284, 295
young workers, 210

Zaller, John, 80
Zegart, Amy, 267
zero-click searches, 55

PROOFS