Expert Insights
PERSPECTIVE ON A TIMELY POLICY ISSUE

BENJAMIN BOUDREAUX, BEBA CIBRALIC

Artificial Intelligence and the Social Contract

Foundations for Social and Economic Policy Under Technological Transformation

August 2025

This publication has completed RAND's research quality-assurance process but was not professionally copyedited.

For more information on this publication, visit **www.rand.org/t/PEA3888-1**.

About RAND

RAND is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest. To learn more about RAND, visit www.rand.org.

Research Integrity

Our mission to help improve policy and decisionmaking through research and analysis is enabled through our core values of quality and objectivity and our unwavering commitment to the highest level of integrity and ethical behavior. To help ensure our research and analysis are rigorous, objective, and nonpartisan, we subject our research publications to a robust and exacting quality-assurance process; avoid both the appearance and reality of financial and other conflicts of interest through staff training, project screening, and a policy of mandatory disclosure; and pursue transparency in our research engagements through our commitment to the open publication of our research findings and recommendations, disclosure of the source of funding of published research, and policies to ensure intellectual independence. For more information, visit www.rand.org/about/research-integrity.

RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Published by the RAND Corporation, Santa Monica, Calif.
© 2025 RAND Corporation
RAND® is a registered trademark.

Limited Print and Electronic Distribution Rights

This publication and trademark(s) contained herein are protected by law. This representation of RAND intellectual property is provided for noncommercial use only. Unauthorized posting of this publication online is prohibited; linking directly to its webpage on rand.org is encouraged. Permission is required from RAND to reproduce, or reuse in another form, any of its research products for commercial purposes. For information on reprint and reuse permissions, visit www.rand.org/about/publishing/permissions.

PE-A3888-1

About This Paper

As the world, and the United States in particular, confronts the challenges of economic mobility, health and longevity, and new frontiers of technology and human well-being, have we designed our social and economic systems to effectively address these convergences? RAND is addressing these questions through an effort called the Social and Economic Policy Rethink to transform the approach to solving social and economic policy problems. The broader volume encompasses seven publications that describe key social and economic policy stakes for policy on artificial intelligence (AI) adoption, examine how policymakers can account for rapid AI capabilities development and adoption in and across sectors, and showcase how industry and policy leaders can responsibly meet the challenges of an era of transformative AI. In this paper, we argue that social contract theory provides a valuable compass for navigating the societal and economic impacts of advances in AI.

RAND Social and Economic Well-Being

RAND Social and Economic Well-Being is a division of RAND that seeks to actively improve the health and social and economic well-being of populations and communities throughout the world.

Funding

Funding for this research was provided by the contributions of the RAND Social and Economic Policy Advisory Board; generous gifts by Frank M. Clark, Michael J. Critelli, William A. Downe, Jihee Kim Huh and Peter Yun Huh, and the Donald M. James Family Foundation; and income from the operation of RAND Social and Economic Well-Being.

Acknowledgments

We thank the entire Social and Economic Policy Rethink team for their insights throughout this project, especially Jonathan Welburn, Jim Mignano, Chandra Garber, Leah Dion, and Anita Chandra. We also greatly appreciate the insightful comments from reviewers Doug Ligor and Salil Gunashekar.

Executive Summary

As AI capabilities advance, and companies like OpenAI explicitly aim to create AI systems that can perform all economically valuable work, the U.S. faces potential disruptions comparable to the Industrial Revolution, requiring adaptations that ensure economic opportunity and societal resilience. This paper argues that social contract theory, already codified in the U.S. Constitution and throughout American institutions, provides a valuable compass for navigating the societal and economic impacts of AI. We draw from several notable social contract theorists (Hobbes, Locke, Rousseau, Rawls, and Allen) to explore six core ideas that can help address contemporary challenges posed by transformative AI:

1. Legitimate political authority is necessary to enable cooperation. AI systems operating across unregulated domains or in novel ways create new collective action problems requiring action from legitimate authorities.

2. Authorities must protect individual rights and liberties. In its own use of AI, the U.S. government must maintain public trust by preserving citizens' rights, including by accountability and oversight, privacy protections, and accessible appeals processes.

3. Legitimacy is derived from democratic deliberation. The widespread use of AI might hinder democratic deliberation by furthering polarization and extremism, but AI might also provide useful platforms for improving deliberation, for example through tools that solicit citizen's input and identify consensus across divided groups.

4. Thinking about all social positions makes society more just and resilient. Considering policy from the perspective of all affected social positions is helpful to moderate the market's tendency to neglect and externalize harms.

5. An overlapping consensus helps promote political and social stability. Despite deep disagreements about AI's future, stakeholders share key views such as that the pace of AI development outstrips government timescales, that the private sector is the leader in AI development, and that AI evaluation and prediction is imperfect. These shared views provide a foundation for adaptive governance approaches and public-private partnerships.

6. Shared participation requires combatting domination and concentrations of power. The possible concentration of AI in a few technology companies and the potential for widespread labor displacement may threaten democratic equality and underscore a need for new approaches for ensuring economic opportunity.

The piece acknowledges the limits of applying theory to the real world. However, we show that social contract ideas offer guidance for developing plans and policies that support the basic foundations of social cooperation. Rather than viewing AI through a purely technical lens, this approach provides a socio-technical analysis necessary to support human and societal flourishing. We hope this piece demonstrates that there is space for productive conversations

across political divides on questions of political legitimacy and social cooperation as America navigates technological transformation.

Contents

About This Paper iii
Executive Summary iv
Artificial Intelligence and the Social Contract: Foundations for Social and Economic Policy Under Technological Transformation 1
I. Introduction: Social Problems Under Technological Transformation 1
II. Social Contract Ideas Applied to AI Governance 3
III. Critiques and Future Directions 16

Artificial Intelligence and the Social Contract: Foundations for Social and Economic Policy Under Technological Transformation

I. Introduction: Social Problems Under Technological Transformation

The rapid advancement of artificial intelligence (AI) is reshaping the landscape of American social and economic policy. Already progress on problems such as housing, healthcare, disaster response, and criminal justice face difficulties including political mistrust and polarization, uncertainty about the impacts of policy interventions, and questions of when individuals or institutions are responsible. As highly capable AI systems stand to transform the economy and broader society—with AI companies like OpenAI explicitly aiming to automate all economically valuable work[1]—we are at an inflection point. These transformations may produce resource abundance and technological solutions that alleviate social ills, or they could worsen underlying social conditions and concentrate economic power in ways that make problems intractable, or perhaps something in between. The scale and speed of this transition will test America's resilience and ability to adapt.

History demonstrates that rapid technological change can lead to major instability—for instance, the Industrial Revolution led to widespread labor displacement, family restructuring, and social unrest.[2] The transition to advanced and widespread AI technologies may be even more disruptive. At a minimum, it will extend existing trends from contemporary social media and digital technologies, such as mental health impacts and polarization, and may cause new social and economic shocks.[3] As capabilities improve and AI is diffused, we need approaches to help maintain economic opportunity, social cohesion, and democratic legitimacy. America will have to address questions such as the scope of AI-driven decision-making, the implications of rapid automation of labor, and the ways in which we can foster human flourishing and social connection in contexts where AI is omnipresent.[4]

---

[1] OpenAI, "OpenAI Charter," https://openai.com/charter/

[2] PBS, "The Labor Wars in the U.S," https://www.pbs.org/wgbh/americanexperience/features/theminewars-labor-wars-us/

[3] See, as examples, Turkle, Sherry, *Alone Together*, Basic Books, 2011, https://www.hachettebookgroup.com/titles/sherry-turkle/alone-together/9780465093656/?lens=basic-books;

Fisher, Max, *The Chaos Machine*, Little, Brown, 2022, https://www.hachettebookgroup.com/titles/max-fisher/the-chaos-machine/9780316703314/?lens=little-brown;

Haidt, Jonathan, *Anxious Generation*, 2024, https://www.anxiousgeneration.com/book

[4] The Trump Administration has indicated human flourishing is a priority for its approach to AI policy and governance. See, for instance, the Executive Order on AI, which states: "This Executive Order establishes the commitment of the United States to sustain and enhance America's dominance in AI to promote human flourishing,

Social contract theory was initially developed to address foundational questions about government and social cooperation that are still relevant today. Concepts from this tradition, like the protection of rights and the importance of consent of the governed, inspired the U.S. Constitution and enshrined in legislation across domains such as housing, employment, and education, among many others. In this way, the social contract tradition is built into American institutions and has shaped the American experience. Initially developed centuries ago, this approach is technology agnostic, providing thinking that is flexible to address the distinct ways AI may diffuse throughout society. Like all applications of theory to the complex real-world, there are limits to applying social contract theory to contemporary AI challenges, but we can find guidance by reconsidering its core tenets. Especially in contexts where Americans are divided, returning to these foundations may enable productive conversations from shared foundations.

At its core, social contract theory emphasizes a simple bargain that Americans' participation in society is reciprocated with safety, security, and economic opportunity. It offers a vision of America as a collaborative project in which each person's well-being is linked to the well-being of others. This interconnected vision of society stands in contrast to purely individualistic views that may struggle in contexts of shared challenges like natural disasters or economic shocks requiring collective resilience and response. Similarly, a social contract approach to AI governance views the key policy questions not through a purely technical lens managed by analyzing AI models or compute. Rather, the approach underscores that the transition to AI requires a broad socio-technical approach to integrate AI in ways that supports well-being and the institutions that help Americans thrive. Governance must go beyond a narrow focus on atomistic individuals or technical AI alignment; instead, it must center the complex ways AI will impact people's work, communities, and relationships. Analysis of these context-specific deployments will be essential, though with the ideas of the basic contract as a guide.

The remainder of this paper is as follows: we review six core ideas from social contract theory—three from classical approaches and three from more contemporary developments—and provide examples of how they apply to AI governance in an American context. We conclude with some thoughts on the way forward. We will focus on the US government and American society as the locus of the social contract. In this sense, we will take nation-states as they are, rather than consider an alternative set of institutions that may perhaps better implement social contract principles. We posit that other political communities may also benefit from a consideration of social contract theory but leave that out of scope for the current analysis.[5]

---

economic competitiveness, and national security." White House, "Removing Barriers to American Leadership in Artificial Intelligence," Presidential Action, January 2025, https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/

[5] There are other important initiatives that are looking to the social contract to address AI issues in a broader global context, see for instance https://www.agisocialcontract.org

II. Social Contract Ideas Applied to AI Governance

Classical social contract theorists and their intellectual followers provide fundamental insights that remain relevant to modern AI governance. The classical theorists laid the intellectual groundwork that undergirds the U.S. Constitution as well as constitutions of many other democracies. These theorists were, however, also deeply biased against a range of social groups; later thinkers further developed social contract theory to integrate a broader set of perspectives, such as those of women, that were previously excluded. Later thinkers also developed the ideas in ways that made them more applicable, and hence more useful, to cotemporary political life. This is a large field and our discussion below is not intended to be exhaustive or detailed. Instead, we raise the key ideas to inspire the development of concrete applications for AI.

Classical Foundations

Idea 1. Legitimate political authority is necessary to enable cooperation

Thomas Hobbes' offers a straightforward explanation for why humans need a shared and stable political order to begin with.[6] In his *Leviathan*, he imagines human life without government, which he calls the "state of nature."[7] In this condition, he argues, human life would be "solitary, poor, nasty, brutish, and short" because constant fear of the more powerful triggers violence and instability, and the lack of a common power to provide security and enforce agreements renders it impossible to trust other individuals or cooperate. In such a state of nature, rational individuals have the incentive to preemptively attack others or break arrangements for personal gain. Even those who would prefer peace must arm themselves and might strike first. This creates a tragic situation where individually rational actions lead to collectively disastrous results—what is now called a "collective action problem." Hobbes argues that to escape this situation, rational individuals would agree (or "covenant") to establish a sovereign power (a "Leviathan") with authority over the use of force to ensure safety and security. Such a Leviathan can protect individuals from each other and create conditions that render cooperation both possible and desirable.

While Hobbes' specific prescription of near-absolute sovereignty is extreme, the core insight—that the legitimate protection of the state enables cooperation—remains crucial. In contemporary contexts, everyday collective action problems (such as traffic, clean air and parks, disaster response, and many others) can only be resolved when trusted institutions foster and enforce cooperation. This foundational analysis helps us understand when legitimate authority is

[6] This paper is focused on ideas inspired from these thinkers and concepts rather than detailed textual interpretation. For instance, Hobbes discusses a "covenant" rather than a contract, so we've taken some liberties to update these terms for contemporary analysis.

[7] Hobbes, Thomas, *Leviathan*, first published 1651.

necessary and when cooperation may break down, such as when enforcement is weak or when benefits of defection are high.

As it pertains to AI, this idea raises the question, *when and how should legitimate authorities develop and enforce rules to ensure benefits and manage risks*?

Addressing AI Agents

One recent technological and market trend is the widespread development of AI agents, which are systems that can plan over timeframes and take independent actions in the world.[8] Companies are developing agents designed to do *anything* a human could do on a computer, including financial trading, mental health service provision, and all kinds of remote work.[9] This may lead to new collective action problems where these agents interact at machine speeds and in novel ways, raising new questions about the risks from the systems, liability for harms, and the ways they may reshape social and economic life.

With respect to other coordination problems such as automobile traffic, governments play a crucial role in establishing rules, including drivers' licenses, seatbelts, traffic signals, speeding tickets, and physical roads. Similarly, governments may have a role in establishing analogous infrastructure for AI agents that includes identity resolution to certify agents and attribute responsibility, safety procedures to minimize unintended harms, and enforcement mechanisms that ensure deployers follow the rules.[10] Even as existing laws apply to the deployment of agents[11]—including those derived from other principal-agent problems—there may be a need for further technical, social, and legal infrastructure that only a legitimate authority could provide. Indeed, AI agents stand to operate at speeds and scale that stretch existing laws and institutional processes, which may underscore the need for new capabilities such as circuit breakers, manual back-ups, and 'kill switches' that ensure humans can maintain appropriate oversight and take action if AI agents malfunction or produce harm.

---

[8] For discussion, see Lazar, Seth. 2024. "Frontier AI Ethics: Anticipating and Evaluating the Societal Impacts of Language Model Agents." *arXiv*. https://doi.org/10.48550/arXiv.2404.06750.

[9] See, for example, Salesforce's introduction of 'Agentforce: The AI Agent Platform': Agentforce: The AI Agent Platform | Salesforce US

[10] For related ideas, see Chan, Alan, Kevin Wei, Sihao Huang, Nitarshan Rajkumar, Elija Perrier, Seth Lazar, Gillian K. Hadfield, and Markus Anderljung. 2025. *Infrastructure for AI Agents*. arXiv preprint arXiv:2501.10114. https://arxiv.org/abs/2501.10114.

[11] For more on legal frameworks for governing AI agents, see Kolt, Noam. 2025. "Governing AI Agents." *Notre Dame Law Review* 101 (forthcoming). https://doi.org/10.2139/ssrn.4772956.

Idea 2. Legitimate political authority must protect individual rights and liberties

John Locke proposes a framework for understanding legitimate political authority and its limits centered on the protection of rights and the role of consent of the governed.[12] Unlike Hobbes, Locke sees the state of nature as generally peaceful but unstable; people have so-called natural rights to life, liberty, and property, but face difficulties in reliably securing these rights without common institutions. Locke argues that political authority must be based on the consent of the governed and limited to protecting these rights.

This creates two legitimacy 'tests' for government authorities: they must be both (1) consented to by citizens, and (2) exercised only within proper limits. When governments exceed these rights-based limits or operate without (tacit or explicit) consent, they lose legitimacy, and the people retain a right of resistance.

Locke's emphasis on rights protection and limited authority could provide a foundation for resilient societies facing technological change. When institutions reliably protect individual rights, citizens can experiment with and adapt to new technologies, enabling innovation while preventing harm. Similarly, through maintaining the commitment to protect basic rights, the political authorities themselves can leverage new technologies to better advance their core security and other missions. This could create a vital adaptive capacity—societies can incorporate beneficial technological changes while ensuring that they are not used in ways that enable domination or that would generate significant (and appropriate) backlash and resistance.

Locke's framework helps us think about questions like: *What considerations should guide government use of AI?*

Government Use of AI

Governments at Federal and local levels have looked to new AI tools to assist with key governance functions. These tools may make elements of governance more cost-effective, reliable, or radically increase the ability of the government to advance its missions in contexts such as homeland security. preventing crime, collecting revenue, or distributing benefits.

However, the ability of government to effectively and rightfully use these tools depends crucially on the public's trust that they will be used in rights-preserving ways. This is not merely a theoretical point; public resistance has, at times, pushed the government to rewind its technology deployments. For example, full-body scanners that displayed images of travelers' bodies were removed by the Department of Homeland Security after the American public

[12] Locke, John. 1689. *Second Treatise of Government*. Edited by C.B. Macpherson. Indianapolis: Hackett Publishing Company, 1980. Note that like these other early thinkers, Locke's texts focused exclusively on men. Some scholars, including Jeremy Waldron (in *God, Locke, and Equality)*, have argued that that 'the people' in the *Second Treatise* ought to be read to include women. Scholars like Diana Coole and Carol Pateman, however, have provided compelling arguments to show that Locke's theory was exclusionary, which we discuss in the last section of this paper.

expressed privacy concerns.[13] This type of public resistance can take myriad forms and ultimately might reduce funding for technology deployments through decisions by Congress, reduce public uptake in cases where public cooperation is essential, and undermine needed operational partnerships across levels of government and with technology developers.[14]

Locke's conception of rights may suggest the importance of establishing how rights are protected in the context of broad government use of AI, including the role and limits of algorithmic decision-making in rights-impacting applications, proper use of personal data, and the right to contest AI decisions when systems fail. In many cases, existing rights protections will apply to AI, but there may be a need for additional clarity and enforcement on *how*.

A Lockean framework might emphasize that AI systems should enhance rather than undermine individual autonomy and consent. For example, when AI is deployed in criminal justice for risk assessments or in healthcare for treatment recommendations, individuals should retain the right to understand the basis of these decisions, as well as having appropriate opt-out options available. Organizations implementing AI in sensitive domains like housing, employment, or lending might be required to demonstrate that their systems do not arbitrarily deprive people of property rights.[15] These approaches need not prevent technological innovation or economic growth. Just as making air travel safer increases the demand among the public to travel, setting standards to ensure rights protections may provide new innovation opportunities that expand rather than contract the AI marketplace.

The Lockean social contract idea also implies that citizens should have transparent mechanisms to petition for redress when AI systems make consequential errors, such as incorrectly denying benefits or flagging innocent individuals as security risks.[16] This might manifest as independent review boards with technical expertise that can audit AI systems when challenged, much as citizens in Locke's framework have the right to appeal to impartial judges when their rights are threatened.[17] All AI deployers, but especially government users of AI, will

---

[13] Boudreaux et al "The Department of Homeland Security's Use of Emerging Technologies: Why Public Perception Matters" https://www.rand.org/pubs/perspectives/PEA691-1.html

[14] Ibid.

[15] Consider some of the examples discussed in Virginia Eubanks *Automating Inequality.* Additionally, see: Angwin, Julia. 2022. "The Seven-Year Struggle to Hold an Out-of-Control Algorithm to Account." *The Markup*. October 8. https://themarkup.org/newsletter/hello-world/the-seven-year-struggle-to-hold-an-out-of-control-algorithm-to-account.

[16] There are an increasing number of examples of government's globally using flawed automated systems. For instance, in the Netherlands the use of an algorithm for tracking benefits fraud that wrongfully punished poor families (see: https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/). Or in Australia, an AI wrongfully accused 400,000 welfare recipients of wrongfully reporting income (see https://www.reuters.com/article/business/media-telecom/feature-australian-robodebt-scandal-shows-the-risk-of-rule-by-algorithm-idUSL8N3340SN/)

[17] In a similar vein, consider this paper on AI contestability: Landau, Susan, James X. Dempsey, Ece Kamar, Steven M. Bellovin, and Robert Pool. 2024. *Challenging the Machine: Contestability in Government AI Systems*. arXiv preprint arXiv:2406.10430. https://arxiv.org/abs/2406.10430.

need to understand when their systems impact rights and the nature of those impacts through regular assessments.

Idea 3. The legitimacy of political authority is derived from democratic deliberation

A third classical theorist, Jean-Jacques Rousseau, offers further insights about democratic decision-making and the relationship between private and public interests.[18] His concept of the "general will" represents an attempt to understand how a democratic society can make legitimate collective decisions while respecting individual freedom. Unlike "private will" (which focuses on narrow self-interest) or the "will of all" (which is just an aggregation of discrete private wills), the general will emerges when citizens deliberate together about the common good. This creates a situation where we follow laws that we give ourselves through a process of collective deliberation.

This process of democratic deliberation involves negotiating *together* to determine the rules to which all agree. In other words, it is this deliberative process which creates the collective of Americans ("We the People") together in a shared political project.

This social contract idea helps consider questions such as, *how does AI affect democratic deliberative processes?*

AI and Democratic Deliberation

Rousseau's insights suggest the development of mechanisms that enable meaningful citizen participation in technology decisions. As AI systems increasingly shape public resources and opportunities, leaving governance to powerful companies or government officials without broader deliberation may create a democratic deficit. Similarly, making assumptions about citizens' interests may miss opportunities to engage the public to facilitate collaboration. Specifically for AI governance, this suggests the importance of creating deliberative forums where citizens can determine together how these technologies should be integrated into social and political institutions, especially when individuals cannot opt out, such as the use of facial recognition technologies in public spaces. Without such deliberative mechanisms, technological deployment risks denying citizens their key role in democratic participation and potentially exacerbating mistrust of governmental institutions and undermining social cohesion.

AI Hindrances: AI may present challenges to meaningful deliberation. The opacity of many AI systems—often characterized as "black boxes"—makes informed public discourse difficult when citizens cannot understand how decisions are made. This technical inscrutability undermines a core premise of deliberation: that participants can meaningfully evaluate and consent to the systems governing them. Further, AI-powered recommendation systems on social media and digital space have accelerated political polarization by creating echo chambers and

[18] Rousseau, Jean-Jacques. 1762. *The Social Contract*. Translated by Maurice Cranston. London: Penguin Books, 1968.

amplifying extreme content, fragmenting the shared information environment necessary for productive deliberation.[19] Some algorithmic systems actively exploit psychological vulnerabilities to maximize users' attention, rather than supporting the reflective and open spirit of democratic deliberation. They can also distort the information ecosystem by promoting sensationalist content that heightens emotions, making public opinion appear more polarized than it actually is. When systems prioritize engagement, a small group of extreme users can crowd out moderate participants, further exacerbating division.[20]

AI Enablers: These challenges to deliberation suggest that achieving the vision of democratic legitimacy in the age of AI requires not only new deliberative forums, but also technical transparency requirements and algorithmic design principles that prioritize democratic values over engagement metrics. Indeed, AI might be assistive to that end. For instance, Taiwan has pioneered a process that combines online and in-person deliberation using the Polis platform, which uses machine learning to identify areas of consensus among competing viewpoints.[21] Unlike traditional forums where the most vocal participants may dominate, Polis visualizes clusters of opinion and highlights statements that achieve broad agreement across otherwise divided groups. This approach could be particularly valuable for AI governance by enabling broader participation while using algorithmic tools to identify areas of overlapping agreement. More generally, collective dialogue systems offer opportunities to create more informative, inclusive, and scalable feedback loops between the public and decision-makers. These systems allow participants to express themselves in their own words, engage in parallel so everyone gets a fair hearing, interact after considering others' contributions, and synthesize the perspectives of participants.[22]

---

[19] Tufekci, Zeynep. 2018. "YouTube's Radicalization Machine." The New York Times, March 10. https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html; Haidt, Jonathan, and Tobias Rose-Stockwell. 2019. "The Dark Psychology of Social Networks." The Atlantic, December. https://www.theatlantic.com/magazine/archive/2019/12/social-media-democracy/600763/; Haidt, Jonathan. 2022. "Why the Past 10 Years of American Life Have Been Uniquely Stupid." The Atlantic, May. https://www.theatlantic.com/magazine/archive/2022/05/social-media-democracy-trust-babel/629369/

[20] See Goldberg, Beth, Diana Acosta-Navas, Michiel Bakker, Ian Beacock, Matt Botvinick, Prateek Buch, Renée DiResta, Nandika Donthi, Nathanael Fast, Ravi Iyer, Zaria Jalan, Andrew Konya, Grace Kwak Danciu, Hélène Landemore, Alice Marwick, Carl Miller, Aviv Ovadya, Emily Saltz, Lisa Schirch, Dalit Shalom, Divya Siddarth, Felix Sieker, Christopher Small, Jonathan Stray, Audrey Tang, Michael Henry Tessler, and Amy Zhang. 2024. AI and the Future of Digital Public Squares. arXiv preprint arXiv:2412.09988. https://arxiv.org/abs/2412.09988.

[21] To take one example of its use, Taiwan sought to regulate ride-sharing services in 2015, and the government used Polis to engage over 4,000 citizens in online deliberation, followed by face-to-face stakeholder meetings. The process revealed unexpected consensus positions—including agreement that taxi drivers and ride-share drivers should follow the same insurance and safety requirements—that formed the foundation for subsequent legislation.

[22] See ideas presented here: Goldberg, Beth, Diana Acosta-Navas, Michiel Bakker, Ian Beacock, Matt Botvinick, Prateek Buch, Renée DiResta, Nandika Donthi, Nathanael Fast, Ravi Iyer, Zaria Jalan, Andrew Konya, Grace Kwak Danciu, Hélène Landemore, Alice Marwick, Carl Miller, Aviv Ovadya, Emily Saltz, Lisa Schirch, Dalit Shalom, Divya Siddarth, Felix Sieker, Christopher Small, Jonathan Stray, Audrey Tang, Michael Henry Tessler, and Amy Zhang. 2024. *AI and the Future of Digital Public Squares*. arXiv preprint arXiv:2412.09988. https://arxiv.org/abs/2412.09988.

There are other approaches to deliberative democracy that may help bolster shared views on complex AI governance questions. For instance, citizen assemblies are attempts to bring together a random group of citizens to reach consensus on policy questions.[23] These have been instituted in a variety of contexts, from Ireland's Constitutional policies regarding abortion[24] to UK approaches to carbon emissions.[25] A growing chorus is making public calls for citizen assemblies for AI,[26] and some attempts have already been made, such as the 2023 People's Panel on AI.[27]

A variety of questions about the use of AI may be posed to these assemblies, especially questions regarding the appropriate role and limits for how government itself should be using the technology, for instance the appropriate uses of facial recognition in public spaces or protests, predictive policing tools, or the use of algorithmic decision-making by government to determine benefits. Many Americans may not be tracking debates about emerging technology or otherwise may remain neutral or apathetic.[28] Alternative deliberative approaches may be needed useful to better inform neutral Americans, build AI literacy, and help them participate in deliberations.

Contemporary extensions

To extend the classical social contract principles, we draw from more contemporary scholarship that has developed classical thinking in ways that reflect America's evolving commitment to equality in ways inclusive of communities, such as women and Black Americans, that were previously excluded.

Idea 4. Thinking about all social positions makes society more just and resilient

John Rawls furthered social contract thinking by introducing the "veil of ignorance" thought experiment, which asks us to evaluate political principles as if we did not know our own position in society.[29] In this thought experiment, we are asked to imagine ourselves in an "original position" where we must choose the basic rules for society without knowing any specifics about

---

[23] Lempert, Robert. 2022. "A Way to Return the Constitution to the People." *Noema Magazine*. July 19. https://www.noemamag.com/a-way-to-return-the-constitution-to-the-people/.

[24] Farrell, D. M., Suiter, J., & Harris, C. (2019). 'Systematizing' constitutional deliberation: the 2016–18 citizens' assembly in Ireland. Irish Political Studies https://www.researchgate.net/publication/328674068_'Systematizing'_constitutional_deliberation_the_2016-18_citizens'_assembly_in_Ireland

[25] Climate Assembly UK. 2020. *The Path to Net Zero*. https://www.climateassembly.uk/report/read/final-report.pdf.

[26] Stilgoe, Jack, "AI Has a Democracy Problem. Citizens' Assemblies Can Help," *Science*, Vol. 385, 2024. DOI: 10.1126/science.adr6713.

[27] Connected by Data, "2023 People's Panel on AI," https://connectedbydata.org/projects/2023-peoples-panel-on-ai

[28] Boudreaux, Benjamin, Douglas Yeung, Rachel Steratore, Thomas Goode, Nidhi Kalra, Sydne J. Newberry, Michael W. Robbins, Natalia Henriquez Sanchez, Keller Scholl, Victoria M. Smith, and Kristin Warren, Public Perceptions of Artificial Intelligence for Homeland Security, Homeland Security Operational Analysis Center operated by the RAND Corporation, RR-A691-1, 2024. As of May 22, 2025: https://www.rand.org/pubs/research_reports/RRA691-1.html

[29] Rawls, John, *A Theory of Justice*, Harvard University Press, 1971.

ourselves—not our race, gender, wealth, natural abilities (including talents), or social status. Behind this veil of ignorance, Rawls argues, rational people would choose principles that protect the most vulnerable positions, since we might end up in any of those positions ourselves. In those conditions, we would want to ensure that even the worst-off in society are still living decent lives. This leads, according to Rawls, to principles that guarantee basic rights and liberties for all, ensure genuine equality of opportunity, and only permit inequalities that work to benefit those who have the least.[30] The power of this approach lies in how it leads us to consider policies from all possible social positions, not just our own perspective, and expands our compassion and empathy for all in our shared political community.

By focusing attention on all positions in society, this framework helps prevent the type of extreme inequality that may destabilize societies; while Rawls' theory is ahistorical, we posit that this framework may prove especially useful during times of change and transition. Technologies like AI could create uneven impacts—potentially concentrating benefits while distributing harms—which can undermine the trust and cooperation necessary for collective adaptation.[31] The veil of ignorance helps correct this tendency by ensuring that technological deployment includes protections for those who might otherwise be excluded, and it may enhance resilience in the context of AI governance by encouraging policies that maintain social cohesion through periods of technological disruption.

This social contract idea motivates questions such as, *how should we consider the broad range of impacts that technology deployments have across society?*

Considering AI's Societal Impacts

The veil of ignorance framework offers a methodology to correct the inherently partial evaluations that characterize technological development in market economies. While technology firms operate under market incentives that reward attention to potential benefits and revenue-generating applications, these same incentives systematically undervalue or obscure potential harms, particularly those manifesting over extended timeframes. Many of the harms of deployments are externalities that many developers do not internalize, and they may accumulate over time in ways that create systemic risks.

By encouraging decision-makers to consider how technology deployments might affect individuals occupying all societal positions, the veil of ignorance facilitates a comprehensive mapping of potential impacts that market mechanisms alone cannot generate. This necessitates a pluralistic evaluation ecosystem where governmental bodies, civil society organizations, academic researchers, and community representatives contribute perspectives. Such multi-stakeholder assessment regimes can illuminate externalities, distributional consequences, and

---

[30] Like with discussion of other rich thinkers, we are glossing over many specific details.

[31] See for instance, Acemoglu, Daron "Harms from AI" https://economics.mit.edu/sites/default/files/2023-07/Harms%20of%20AI.pdf Also published in Bullock, Justin *The Oxford Handbook of AI Governance*

second-order effects that remain invisible when technologies are evaluated solely through the lens of their developers' interests, ultimately producing more accurate benefit-cost analyses that incorporate the full spectrum of societal impacts.

Recent research by Weidinger et al. demonstrates how this framework can be practically applied to AI governance.[32] Their study shows that when participants chose principles to govern an AI assistant from behind a veil of ignorance—without knowing their own position in a group—they consistently preferred principles that prioritized helping the worst-off members rather than principles that maximized overall productivity. Weidinger et al.'s research also shows that preferences formed behind the veil of ignorance tend to be more stable and resistant to self-interested revision, suggesting that governance frameworks developed through this approach may maintain broader legitimacy and resilience through periods of technological change.

This research offers guidance for applying the veil of ignorance to AI governance in several areas. For instance, in automated hiring systems, policymakers could use a veil of ignorance exercise where they design regulations without knowing whether they would be job applicants with different backgrounds, employers needing specific talents, or regulators concerned with systemic outcomes. This approach may highlight considerations for transparency of assessment criteria and meaningful appeals processes. Or to consider another example in healthcare resource allocation algorithms, the veil of ignorance would ask decision-makers to establish principles without knowing whether they would be patients with rare or common conditions, physicians with various specialties, or administrators managing limited budgets. A veil of ignorance approach may lead to systems that establish a floor of guaranteed care and could identify ways AI may assist with that goal.

Idea 5. An overlapping consensus helps promote political and social stability

Rawls's additional concepts of overlapping consensus and public reason provide guidance for how democratic societies can achieve stable cooperation despite deep disagreements among citizens.[33] An overlapping consensus occurs when people who hold different religious, moral, or philosophical doctrines can nevertheless agree on basic political principles for structuring society. For example, citizens might have different religious views about the ultimate meaning of life but still agree on principles of religious tolerance and separation of church and state. "Public reason" refers to how we should justify policies in a pluralistic democracy—by appealing to principles and evidence that all reasonable citizens could accept, regardless of their particular

---

[32] Weidinger, Laura, Kevin R. McKee, Richard Everett, Saffron Huang, Tina O. Zhu, Martin J. Chadwick, Christopher Summerfield, and Iason Gabriel. 2023. "Using the Veil of Ignorance to Align AI Systems with Principles of Justice." Proceedings of the National Academy of Sciences of the United States of America 120 (18): e2213709120. https://doi.org/10.1073/pnas.2213709120.

[33] Rawls, John, *Political Liberalism,* Columbia University Press, 1993; Rawls, John, "The Idea of Public Reason Revisited," *University of Chicago Law Review*, Vol. 64, No. 3, Summer 1997.

comprehensive perspective.[34] This means focusing on shared political values and publicly accessible evidence rather than deeply contested religious or philosophical claims.

While there are conceptual arguments against Rawls' framing—for example, the political principles Rawls takes to be neutral may in fact be expressions of particular doctrines—these ideas nevertheless inspire social resilience and cohesion. When diverse citizens can identify shared values despite their differences, societies can adapt to technological change without fragmenting along ideological lines.

This social contract idea motivates questions such as, *even in a context of disagreement and uncertainty, what are some shared or consensus ideas about AI?*

Shared Understanding of AI Challenges

There is significant uncertainty about AI, but there may be significant value in starting from shared understanding of the challenges. This becomes particularly relevant in current AI policy debates, where stakeholders with vastly different worldviews—from accelerationists to doomers, from market-libertarians to social-progressives—must find common ground for governance. Fortunately, there appear to be some basic agreements about the technology and how it will diffuse. To note a few examples:

Pace of Development: Americans across the spectrum may broadly agree that the pace of AI development and diffusion outstrips the timescales on which democratic institutions typically operate. While major legislation often takes years to develop, AI tools have been created and adopted[35] across society within months. This temporal mismatch creates governance gaps where powerful technologies deploy before democratic processes can act.

One possibility based in this shared challenge is to explore more rapid approaches to implement policy, such as adaptive governance approaches and regulatory sandboxes. These approaches enable experimentation with governance mechanisms that can evolve alongside the technology, rather than trying to craft perfect regulations upfront. Regulatory sandboxes create controlled environments where innovations can be tested with appropriate safeguards while simultaneously generating evidence for more permanent governance structures.[36] Additionally,

---

[34] For Rawls, "reasonable" implies a willingness to propose and abide by fair terms of social cooperation, even when those terms might not perfectly align with one's own deeply held comprehensive moral, religious, or philosophical beliefs, recognizing that others, too, hold diverse yet equally valid perspectives. For those interested in further study of Rawls, we recommend examining the concept of "reflective equilibrium", the internal reasoning process that helps individual examine and justify their own beliefs.

[35] This is not to suggest that all AI development and diffusion happens within months. As many researchers have argued, including those most concerned with rapid development timelines, it can take years to develop AI capabilities and even longer for those developments to be diffused across different economic and social sectors.

[36] For discussion about regulatory sandboxes in the context of AI, see Perault, Matt, "Setting the Agenda for Global AI Leadership: Assessing the Roles of Congress and the States," https://a16z.com/setting-the-agenda-for-global-ai-leadership-assessing-the-roles-of-congress-and-the-states/

governance bodies with sufficient technical expertise and streamlined processes can respond more nimbly than traditional legislative cycles.

Public/Private Balance: A second element of overlapping consensus is that the private sector is the leader in AI development and deployment, and has more power in shaping the trajectory of AI than in other cases involving emerging and novel technologies. These private actors include frontier AI companies developing new models, legacy technology firms with the infrastructure and computing power to apply AI across a range of use-cases, and various businesses in almost all sectors—including in high-risk ones such as finance and healthcare—that are racing to deploy AI before their competitors beat them to it. The government, meanwhile, is challenged with determined how to manage potential safety and societal risks without hindering innovation. A range of factors further exacerbate its difficulties, including a lack of technical expertise in government, disagreement over which kinds of regulation are needed, and growing international competition.

One option is to develop institutional arrangements that acknowledge private sector innovation while establishing appropriate public oversight, particularly for high-risk applications affecting fundamental rights and essential services. Public-private partnerships with clearly defined accountability mechanisms can leverage market dynamism while maintaining democratic legitimacy. International coordination mechanisms—whether formal treaties or informal standards alignment—can address global jurisdictional challenges by creating consistent governance expectations across borders. Additionally, placing stronger conditions on government procurement of AI technologies can use public purchasing power to shape private sector development practices, establishing democratically determined guardrails without directly regulating innovation.

Complexity of Evaluations: A third element of overlapping consensus is that evaluating and benchmarking AI—meaning how to assess the actual capabilities of advanced AI models—is inherently complex, and we need to develop a more robust scientific approach. As it stands, the speed and emergence of AI capabilities is outpacing the development of relevant and comprehensive evaluations. Further, ensuring that benchmarks reflect real-world complexities and are not just gamed by developers optimizing specifically for those metrics poses also poses difficulties.

Public reason offers valuable guidance amid intense disagreement about the extent these systems pose catastrophic risks. The US's AI Safety Institute, the UK's AI Security Institute, and the corresponding international network exemplify this approach by establishing evaluation protocols justified through shared public values around preventing harm rather than contested claims about superintelligence, transhumanism, or other speculative concepts.[37] By focusing on broadly accepted concerns—such as preventing deception and bioweapon or cyber-attacks—

[37] See for instance work by the UK Security Institute on making evaluations more actionable: https://www.aisi.gov.uk/work/making-safeguard-evaluations-actionable

these institutions create security frameworks that can build support from those with different views on AI's long-term trajectory.

This framework extends to both laboratory evaluations commonly conducted by AI developers and socio-technical field testing. In laboratory settings, evaluations can assess capabilities, vulnerabilities, and potential harms in controlled environments experiments and exercises that are designed to identify risks. These technical evaluations create a baseline understanding of AI system behaviors that diverse stakeholders—from industry developers to safety researchers—can observe regardless of their ultimate views on AI governance.

Complementary socio-technical field evaluations, which are currently under-resourced, examine how AI systems function in real-world contexts and assess downstream societal impacts that may not be apparent in the lab. For example, field tests of hiring algorithms across different demographic communities could reveal disparate impacts that controlled evaluations might miss. Similarly, socio-technical evaluations would help assess outcomes from the deployments of mental health AI or other highly engaging AI systems. These field evaluations create an empirical foundation for governance conversations by shifting discussions from contested theoretical claims about AI's future to observable evidence about concrete harms and benefits.[38] When combined, lab and field evaluations create an evaluation ecosystem that can adapt as technologies evolve. This adaptability is essential for maintaining overlapping consensus through periods of rapid technological change, as it allows governance frameworks to incorporate new evidence without requiring continuous renegotiation of fundamental principles.

Idea 6. Shared participation requires combatting domination and concentrations of power

Lastly, Danielle Allen's work on democratic equality and the social contract emphasizes how sharing a political community requires attention to both participation and power, with a particular focus on preventing domination in all its forms. Allen's work is focused on democracy and she argues that democracy's fundamental aim is to prevent domination—whether from government, private power, or social structure—where domination occurs when some individuals or groups can exercise arbitrary power over others without accountability or justification.[39] For Allen, avoiding domination requires more than formal guarantees of rights or periodic voting in elections. Instead, it demands economic security and other prerequisites for active participation in collective decision-making. This helps explain why policies concerning education, labor protections, and economic opportunity are not just optional benefits but fundamental requirements for effective democratic participation.

---

[38] Organizations like the Algorithmic Justice League have pioneered field auditing methodologies that document real-world AI impacts in domains like facial recognition. For more, see: https://www.ajl.org

[39] See Allen, Danielle S., *Our Declaration: A Reading of the Declaration of Independence in Defense of Equality*, Liveright, 2014; Allen, Danielle S., *Justice by Means of Democracy*, University of Chicago Press, 2023.

Allen emphasizes how significant economic inequality can create forms of domination that undermine democratic citizenship. When some citizens lack basic economic security or face structural barriers to opportunity, they become vulnerable to unaccountable public and private decisions. This suggests that extreme inequalities of wealth or power are incompatible with genuine democracy, not just because they are unfair, but because they enable forms of domination that make equal citizenship impossible. This has direct implications for policy design such as economic concentration or how to mitigate the harms that will come with labor displacement.

This social contract idea motivates questions such as, *how can we ensure democratic equality in contexts where human labor may be widely displaced by AI?*

Power Concentration and Labor Displacement

The concentrated ownership of frontier systems by a few technology companies creates risks of domination when the firms can unilaterally shape information environments and deploy powerful tools that affect millions. The ability of AI companies to leverage vast datasets into continually improving systems creates forms of market power perhaps fundamentally different from historical monopolies in physical industries. Similarly, as human labor is replaced by capital automation, the owners of capital stand to significantly gain relative to workers, and where companies need fewer workers to make productivity gains.

Allen's framework would prioritize institutional arrangements that give affected communities meaningful voice in governance decisions. As examples, this may occur through multistakeholder oversight boards with genuine decision-making authority, or through requirements that deployment decisions for high-impact systems undergo public interest reviews that consider risks.

Allen's framework may also offer valuable guidance for addressing wealth concentration and labor displacement due to automation. Automation technologies that replace labor without creating new roles for humans can lead to severe economic insecurity, as well as concentrated capital ownership.[40] When companies deploy warehouse automation while capturing most productivity gains, they create asymmetric power relationships. This suggests the need for forms of economic rebalancing to ensure that displaced workers can still flourish with new opportunities or economic support, although the precise ways we could do this requires further research and experimentation. Allen’s work can be usefully extended to help us critically examine the power structures that underlie the institutions, both public and private, that wield most power in AI development and deployment.

---

[40] See, for example, Acemoglu, Daron, and Simon Johnson, Power and Progress, PublicAffairs, 2024, https://www.hachettebookgroup.com/titles/daron-acemoglu/power-and-progress/9781541702530/?lens=publicaffairs

III. Critiques and Future Directions

The social contract tradition has faced critiques from a variety of political perspectives. While we suggest the theory has broad and deep applicability to AI challenges, understanding its limits allows us to apply it more conscientiously.

One set of critiques concerns the abstract nature of social contract theory. Iris Marion Young argues that social contract theory's abstraction from real historical conditions masks how power relations shape both the development and implementation of social arrangements.[41] Young contends that the theory's emphasis on impartial rules and universal principles obscures how seemingly neutral institutions can reproduce systematic inequalities. The very idea of autonomous individuals freely entering into social contracts ignores how identities and interests are shaped by existing social relations and power structures. Others, such as Carole Pateman[42] and Charles Mills,[43] argue that historical social contracts have systematically excluded women and racial minorities, making the theory's claims to universality deeply problematic. Their work suggests that rather than representing genuine universal agreement, social contracts often codify existing privileges and exclusions. Their critiques become particularly relevant when considering how new technologies might embed and amplify existing patterns of exclusion. (Contemporary scholars have made an analogous argument about AI and automated systems, suggesting that they encode and hide existing privileges.[44])

Another cluster of critiques focuses on the validity of concepts encompassed in social contract theory. For example, libertarian thinker Robert Nozick argues that social contract theory often serves to justify more extensive state intervention than could be supported by genuine voluntary agreement.[45] Another critique is from Austrian economist Friedrich Hayek, who questions whether any central authority can have sufficient knowledge to design social institutions in the way social contract theory suggests. He argues that social order emerges

---

[41] Young, Iris Marion, *Justice and the Politics of Difference*, Princeton University Press, 1990.

[42] See, for example, Pateman, Carole, *The Sexual Contract*, Stanford University Press, 1988.

[43] See, for example, Mills, Charles W., *The Racial Contract*, Cornell University Press, 1997.

[44] While these scholars may not critique social contract theory specifically, the structure of their critique of AI aligns with that of Mills and Pateman of social contract theory, respectfully. For example, Ruha Benjamin's analysis reveals how technological design choices that appear neutral often encode and amplify racial hierarchies (e.g. *Race After Technology, the New Jim Code)*. Joy Buolamwini's research demonstrates how facial recognition systems exhibit systematic bias against darker-skinned faces, particularly women of color, highlighting how seemingly objective technologies can reproduce historical patterns of exclusion (e.g. *Gender Shades, Unmasking AI)*. And Timnit Gebru's work shows how large language models can amplify societal biases while obscuring their training data and development processes from democratic oversight (e.g. Gebru, T., et al." On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" ACM Conference on Fairness, Accountability, and Transparency 2021).

[45] See, for example, Nozick, Robert, *Anarchy, State, and Utopia*, Basic Books, 1974.

through spontaneous evolution than through rational design, making social contract theory's attempt to derive principles for institutional design fundamentally misguided.[46]

These critiques highlight important limitations of social contract approaches. They suggest the need to supplement social contract analysis with careful attention to power relations, historical conditions, and the practical limitations of institutional design. However, rather than completely invalidating social contract theory, these critiques can help refine its application to contemporary challenges, particularly as we navigate the transformational changes brought by AI.

Fortunately, we are not starting from scratch; many social contract principles found concrete expression in America's founding documents and institutional design. The Constitution's system of checks and balances reflects Lockean concerns about limiting government power, while its federal structure attempts to balance local and national interests in ways that apply Rousseau's thinking about collective decision-making. The Bill of Rights codifies citizens' rights, with its protections of property rights, religious freedom, and free speech. And the Commerce Clause reflects a Hobbesian understanding that national-level coordination is sometimes necessary to overcome collective action problems between states. Later developments in American law, from antitrust legislation to civil rights protections, continue to engage with these fundamental questions about balancing individual liberty with collective goods and establishing legitimate authority with clearly demarcated scope.

Looking ahead, the trajectory of AI suggests a multitude of domestic American challenges, including **economic disruption** from widespread automation; **democratic degradation** as AI may undermine shared epistemic foundations; **power concentration** among technology companies who control AI, creating new forms of authority beyond democratic accountability; **weakening human agency and social cohesion** with psychological and social effects from increased human dependency on AI; and **institutional gaps** as AI technologies evolve faster than institutional responses.

Addressing these challenges may requires capabilities and plans that do not yet exist. The social contract idea offers a compass (rather than a blueprint) to navigate technological transition, and helps indicate the direction and possible measures for human and societal flourishing. The task ahead requires translating enduring and widely endorsed principles into effective institutional capabilities and plans, ensuring technological progress serves rather than subverts our shared American society.

---

[46] See especially Hayek's *Law, Legislation, and Liberty* https://www.routledge.com/Law-Legislation-and-Liberty-A-New-Statement-of-the-Liberal-Principles/Hayek-Shearmur/p/book/9780415035323?srsltid=AfmBOooqWPKpgpTsfEJ0h5fWYVdTTY09DcGoa7_x7QJlmMFC3S3lvQE8