ARTIFICIAL INTELLIGENCE AND DEMOCRATIC FREEDOMS



April 10-11, 2025

KNIGHT
FIRST AMENDMENT
INSTITUTE at
COLUMBIA UNIVERSITY

Welcome 2

Introduction 3

Program at a Glance 4

Program

- Panel 1: Regulating AI in a Time of Democratic Upheaval 9
- Panel 2: AI Agents' Democratic and Economic Impacts 13
- Keynote Address: Ben Buchanan 17
- Panel 3: Evaluation and Design of Safe AI 19
- Panel 4: Using AI to Enhance Democracy 23
- Panel 5: Reframing AI in the Digital Public Sphere 27
- Panel 6: AI, Freedom of Expression, and Democratic Backsliding 31
- Keynote Address: Erica Chenoweth 33
- Panel 7: Transitions to Transformative AI 35

Speakers 36

About the Knight First Amendment Institute 48

WELCOME

This symposium takes place at a moment of dramatic political and technological upheaval. Rapid advancements in the capabilities of general purpose artificial intelligence (GPAI) systems have led to predictions of both unprecedented human flourishing and existential threats to humanity. These advancements are taking place against a backdrop of a fracturing postwar world order and an uncertain future for U.S. democracy, raising many questions about how regulation and governance of these systems will be carried out—and with whose interests in mind.

As GPAI systems gain new capabilities and are further integrated into our economy, governance, and communications networks, it is essential to investigate how these systems will challenge democracy, how they may shape—and strain—our public discourse and our democratic institutions, how they may accelerate democratic backsliding, and how they may impact the public's ability to exercise effective self-government.

We must identify strategies to meet these challenges and ensure that these systems serve all of humanity and not just a powerful few. This symposium on artificial intelligence and democratic freedoms, the product of a year-long collaboration with the Knight Institute's Senior AI Advisor Seth Lazar, aims to do this by convening an interdisciplinary group of scholars, advocates, and tech industry professionals. We are deeply grateful to them all for their contributions to this project, and we look forward to a series of engaging and productive conversations over the next two days. The papers they are presenting here will be published on the Institute's website in the coming months.

Finally, we would like to express our appreciation for the Knight Institute's funders (listed at the end of the program), to the Institute staff whose hard work made this symposium possible, and to all of you for joining us.

Jameel Jaffer and Katy Glenn Bass

INTRODUCTION

Artificial intelligence capabilities are rapidly accelerating. No longer constrained to narrow domains, AI is now genuinely general purpose. Beyond generating content in practically any modality, large language models (LLMs) can now provide cognitive resources for capable autonomous systems. In addition to these software agents, there have been striking advances in humanoid robotics in the last year alone. But this rapid progress in science and engineering has been met by maladapted regulatory processes and narrow technical safety interventions that, while essential to bring these systems to market, leave their most consequential societal impacts untouched.

In times that portend radical social change, we might reasonably hope that democratic institutions would provide guardrails and goals. But democracies worldwide are either enfeebled or aggressively backsliding from core democratic values, with commensurate impacts on basic freedoms such as those of expression and thought.

Charting a safe course for society through these twin challenges will require much more than just technical interventions or repurposing legislation prepared for a different era of AI progress. One necessary—though not sufficient—ingredient is to foster a cohort of technical researchers with a deep understanding of the broader sociotechnical systems in which AI is embedded and sociotechnical researchers who are equally familiar with the capabilities and affordances of the AI systems that will undergird the future. This symposium aims to do just this, focusing the collective attention of a methodologically diverse group of scholars on the prospective impacts of AI on democratic freedoms.

Seth Lazar

PROGRAM AT A GLANCE

April 10, 2025

9:30 am – 9:45 am — **Welcome**

Jameel Jaffer

9:45 am – 10:55 am — **Panel 1: Regulating AI in a Time of Democratic Upheaval**

Panelists	Moderator
Atoosa Kasirzadeh Deirdre K. Mulligan Arvind Narayanan Alondra Nelson	Shayne Longpre

10:55 am – 11:10 am — **Break**

11:10 am – 12:20 pm — **Panel 2: AI Agents' Democratic and Economic Impacts**

Panelists	Moderator
Kevin Feng Peter Henderson Seth Lazar Daniel Susskind	Beba Cibralic

12:20 pm – 1:00 pm — **Lunch**

Lunch will be provided

1:00 pm – 2:00 pm — **Keynote Address**

Ben Buchanan

2:00 pm – 2:15 pm — **Break**

2:15 pm – 3:25 pm — **Panel 3: Evaluation and Design of Safe AI**

Panelists	Moderator
Gillian Hadfield Hoda Heidari Lujain Ibrahim Sydney Levine	Helen Toner

3:25 pm – 3:40 pm	**Break**	
3:40 pm – 4:50 pm	**Panel 4: Using AI to Enhance Democracy**	
	Panelists Hahrie Han Spencer Overton MH Tessler Luke Thorburn	**Moderator** Jakob Mökander
4:50 pm – 5:00 pm	**Closing Remarks**	



April 11, 2025

9:30 am – 9:45 am	**Welcome**	
	George Deodatis	
9:45 am – 10:55 am	**Panel 5: Reframing AI in the Digital Public Sphere**	
	Panelists Sacha Altay Joshua P. Darr Manon Revel Ranjit Singh	**Moderator** Nadine Farid Johnson
10:55 am – 11:10 am	**Break**	
11:10 am – 12:20 pm	**Panel 6: AI, Freedom of Expression, and Democratic Backsliding**	
	Panelists Becca Branum Aziz Huq Xiangnong (George) Wang	**Moderator** Katy Glenn Bass
12:20 pm – 1:00 pm	**Lunch**	
	Lunch will be provided	
1:00 pm – 2:00 pm	**Keynote Address**	
	Erica Chenoweth	
2:00 pm – 2:15 pm	**Break**	
2:15 pm – 3:25 pm	**Panel 7: Transitions to Transformative AI**	
	Panelists Justin Bullock Rumman Chowdhury Séb Krier Sam Manning	**Moderator** Seth Lazar
3:25 pm – 3:30 pm	**Closing Remarks**	



PANEL 1 - THURSDAY

Regulating AI in a Time Of Democratic Upheaval

9:45 am - 10:55 am

Liberal democracies had almost reached consensus on how to regulate narrow AI when ChatGPT came along and general purpose AI took center stage. How should regulatory proposals adapt to this new reality—and what prospects are there for effective regulation when the global geopolitical order is being reconfigured?

Panelists
Atoosa Kasirzadeh, *Carnegie Mellon University*
Deirdre K. Mulligan, *University of California, Berkeley*
Arvind Narayanan, *Princeton University*
Alondra Nelson, *Institute for Advanced Study*

Moderator
Shayne Longpre, *Massachusetts Institute of Technology*

PAPER ABSTRACTS

Sociotechnical Challenges in AI Catastrophic Risk Governance
Atoosa Kasirzadeh, *Carnegie Mellon University*

Scaling policies have emerged as critical tools for managing catastrophic AI risks that guide decisions about safely increasing AI model capabilities. This paper argues for grounding these frameworks in sociotechnical considerations. It proposes three approaches to advance AI safety that combine rigorous technical assessment with participatory processes and democratic ideals to develop a socially responsible approach to AI governance while addressing potential systemic impacts.

AI Safety is Sometimes a Model Property: A Conceptual Model to Guide AI Risk Governance Strategy
Deirdre K. Mulligan, *University of California, Berkeley*
Nik Marda, *Mozilla*

This paper examines how safety interventions in AI models might mitigate the broader risks that artificial intelligence poses to the public's rights and safety. It assesses the types of risks that model-level interventions can reasonably address and asks: What capabilities and competencies must an actor possess to effectively and legitimately carry out those interventions? The authors present a conceptual framework for choosing the optimal unit(s) of intervention—i.e., model, system, or sociotechnical system—for risk mitigation and the actors able to legitimately perform the mitigations and for distinguishing between the hazards that AI models may increase and the *harms* they produce.

AI as Normal Technology and Implications for AI Safety

Arvind Narayanan, *Princeton University*
Sayash Kapoor, *Princeton University*

This paper articulates a vision of artificial intelligence as "normal technology," in contrast to both utopian and dystopian visions that treat AI as a potentially superintelligent entity. It explains why the impacts of advanced AI, even if transformative, will unfold slowly, making a critical distinction between AI methods, AI applications, and AI adoption. The paper also discusses a potential division of labor between people and AI in a world with advanced AI and examines the implications of AI as normal technology for AI policy, AI safety, and human progress.



PANEL 2

AI Agents' Democratic and Economic Impacts

11:10 am - 12:20 pm

The most consequential societal impacts of large language models will come not through the content they generate but through the cognitive resources they provide to effective software agents. This panel considers how these agents in turn will reconfigure political and economic realities.

Panelists

Kevin Feng, *University of Washington*
Peter Henderson, *Princeton University*
Seth Lazar, *Australian National University and the Knight First Amendment Institute*
Daniel Susskind, *King's College London*

Moderator

Beba Cibralic, *RAND*

PAPER ABSTRACTS

Levels of Autonomy for AI Agents
Kevin Feng, *University of Washington*
David W. McDonald, *University of Washington*
Amy X. Zhang, *University of Washington*

The excitement over autonomous AI agents is driving their imminent inclusion into society, but simultaneously, AI's increasing autonomy raises significant safety concerns. How can agent designers better reason about the appropriate levels of autonomy at which their agents should operate? This paper introduces a five-level framework for agent autonomy, advocating for autonomy as a deliberate design decision that need not be inherently tied to agent capability. The authors discuss potential applications and implications of this framework for both single- and multi-agent systems. This work aims to take meaningful, practical steps towards safe and useful AI agents in the real world.

AI Agents and Control
Peter Henderson, *Princeton University*

As we shift from static AI systems toward agentic systems that can operate autonomously in the world, AI will be able to supercharge influence operations. To mitigate this risk, AI model creators already train models to avoid persuasion and manipulation, monitor their use, and apply measures like watermarking to AI-generated content. This paper argues that trying to detect agent-powered influence operations at the model level will not always succeed and proposes potential technical alternatives—such as using AI agents to disrupt other, harmful agents in the wild—that move beyond the typical safety framework.

Agents For and Against Democracy

Seth Lazar, *Australian National University and the Knight First Amendment Institute*
Mariano-Florentino Cuéllar, *Carnegie Endowment for International Peace*

Each advance in computational technology has prompted anxious reflection on its potential impacts on democracy. In 1984, Langdon Winner criticized both optimistic "computer romantics," who envisioned the information age as ushering in an era of participatory and egalitarian democracy, and pessimists, who viewed computational tools as inherently authoritarian. Today, rapid advances in artificial intelligence introduce highly capable and substantially autonomous AI agents that could rapidly proliferate across societal and political domains. This paper investigates how these agents might amplify existing anti-democratic trends, identifies new democratic threats they may uniquely pose, and proposes pathways by which AI agents could instead actively reinforce and defend democratic freedoms.

What Will Remain for People to Do?

Daniel Susskind, *King's College London*

What work might remain for human beings to do if we approach a world where AI is able to perform all economically useful tasks more productively? This paper argues that the common answer to this question—"nothing"—is mistaken. There are good reasons to believe that there will remain tasks for people to do, due to three limits: general equilibrium limits, involving tasks in which labor has the comparative advantage over machines; preference limits, involving tasks where humans might prefer an un-automated process; and moral limits, involving tasks with a normative character, which people might believe require a "human in the loop" to exercise their moral judgment.

KEYNOTE ADDRESS

Ben Buchanan

1:00 pm - 2:00 pm

Ben Buchanan is an assistant professor at Johns Hopkins University's School of Advanced International Studies. He served in the Biden White House for all four years, including as the White House special advisor for AI and as the director for technology and national security on the National Security Council. He is the author of three books on AI, national security, and cybersecurity. Prior to joining the White House, he was a professor at Georgetown University. He earned undergraduate and graduate degrees from Georgetown University and a Ph.D. from King's College London, where he was a Marshall Scholar.



PANEL 3

Evaluation and Design of Safe AI

2:15 pm - 3:25 pm

Advances in large language models were enabled by, and fostered, a boom in technical AI safety research. But narrowly technical approaches not only mischaracterize AI risks, they miss key opportunities for mitigating them. This panel fosters a more multidisciplinary, sociotechnical approach to technical AI safety.

Panelists
Gillian Hadfield*, Johns Hopkins University*
Hoda Heidari*, Carnegie Mellon University*
Lujain Ibrahim*, University of Oxford*
Sydney Levine*, Google DeepMind*

Moderator
Helen Toner*, Georgetown University*

PAPER ABSTRACTS

Building AI for the Democratic Matrix: Normative Competence and Normative Institutions

Rakshit Trivedi, *Massachusetts Institute of Technology*
Gillian Hadfield, *Johns Hopkins University*
Dylan Hadfield-Menell, *Massachusetts Institute of Technology*

Democracy is ultimately an alignment process. It aligns individual behaviors with the collective welfare of a community, as defined by its normative institutions. This paper argues that AI must be aligned not with individual preferences, laws, or norms, but rather with the complex processes that generate democracies as our shared *normative social order*. This requires that AI be built to possess *normative competence* and incentivized to participate in normative behaviors meaning both compliance and distributed enforcement of rules and norms. Additionally, *normative classification institutions* that are legible to AI and operate at the speed and scale of AI are needed. These challenges are grounded in theory of how human normative social orders operate, but they raise fundamentally technical challenges for AI researchers and developers.

The AI Power Disparity Index: Toward a Compound Measure of AI Actors' Power to Shape the AI Ecosystem

Hoda Heidari, *Carnegie Mellon University*
Rachel M. Kim, *Carnegie Mellon University*
Blaine Kuehnert, *Carnegie Mellon University*

What can be done to address the growing power imbalances created by the rise of AI? This paper introduces a new tool to assess and respond to such power imbalances: The AI Power Disparity Index is a composite indicator designed to quantify and signal the changing distribution of power in the AI ecosystem. Drawing on interdisciplinary theories and existing composite measures, the index highlights several dimensions of power, including resource-based measures, decision-making capabilities, agenda-setting power, and behavioral and cultural influence.

Interactive Evaluations for Interaction Harms in Human-AI Systems

Lujain Ibrahim, *University of Oxford*
Saffron Huang, *Collective Intelligence Project*
Lama Ahmad, *Open AI*
Markus Anderljung, *Centre for Governance of AI*
Umang Bhatt, *NYU Center for Data Science, The Alan Turing Institute*

Model evaluations have become central to ensuring AI system safety, but current approaches fail to capture the risks that emerge through sustained human-AI interaction. While existing evaluations can detect issues in individual model outputs, they miss more subtle interaction harms that develop as AI systems influence human relationships, perceptions, and behaviors over time. This paper integrates interdisciplinary insights to propose approaches for evaluating interaction harms. The authors outline principles for designing evaluations that better reflect real-world AI use, measuring changes in human cognition and behavior, and appropriately incorporating human participation in the assessment process.

Using Resource Rational Contractualism for AI Alignment

Sydney Levine, *Google DeepMind*
Seth Lazar, *Australian National University and the Knight First Amendment Institute*
Joshua Tenenbaum, *Massachusetts Institute of Technology*
Yejin Choi, *NVIDIA, University of Washington, Stanford University*
Tan-Zhi Xuan, *Massachusetts Institute of Technology*

How should we align AI systems with human values? Traditional value alignment approaches are poorly suited to handle emerging AI systems that will confront a more varied and less predictable set of stimuli than ever before. This paper argues that resource rational contractualism (RCC) should guide AI alignment. Under this principle, AI systems are guided by simple, articulable principles, supported by a thick reason and sourced through democratic processes and institutions or newly developed strategies for personalized and participatory design. RCC enables many of the characteristics of trustworthy AI, including robustness, explainability, auditability, and generalization, and provides an avenue for value pluralistic alignment.



PANEL 4

Using AI to Enhance Democracy

3:40 pm - 4:50 pm

Democratic freedoms are imperiled by affective polarization and radicalization. This panel explores how AI can, and cannot, be used to forge agreement or even consensus in formal and informal democratic deliberation.

Panelists

Hahrie Han, *Johns Hopkins University*
Spencer Overton, *George Washington University*
MH Tessler, *Google DeepMind*
Luke Thorburn, *King's College London*

Moderator

Jakob Mökander, *Tony Blair Institute for Global Change*

PAPER ABSTRACTS

Can AI Mediation Improve Democratic Deliberation? A Critical Analysis of the Habermas Machine

MH Tessler, *Google DeepMind*
Georgina Evans, *Google DeepMind*
Michiel Bakker, *Google DeepMind*
Iason Gabriel, *Google DeepMind*
Sophie Bridgers, *Google DeepMind*
Rishub Jain, *Google DeepMind*
Raphael Koster, *Google DeepMind*
Verena Rieser, *Google DeepMind*
Anca Dragan, *Google DeepMind*
Matthew Botvinick, *Google DeepMind*
Christopher Summerfield, *University of Oxford*

The strength of democracy lies in the free and equal exchange of views from a variety of perspectives. This paper asks whether and how artificial intelligence could enhance democratic deliberation, through the example of an AI system designed to help people with diverse viewpoints find common ground—the Habermas Machine. The authors raise new questions about the role of AI mediation in the democratic process, how such systems could be synergistic with face-to-face deliberation, and outstanding technical considerations regarding fairness, transparency, inclusion, and participation. Ultimately, a range of empirical, technical, and theoretical advancements are needed to fully realize the promise of AI-mediated deliberation for enhancing citizen engagement.

AI and Democratic Publics

Henry Farrell, *Johns Hopkins University*
Hahrie Han, *Johns Hopkins University*

This paper seeks to reveal the underlying paradigms that guide much of the existing scholarship on AI and democracy and articulate possible alternative approaches that more accurately reflect both empirical and normative descriptions of a healthy democracy. The authors consider the relationship between AI and democracy as one that is focused on the construction of democratic publics that shape collective behaviors and ask: How might AI reshape democratic publics? Will these newly formed publics be more or less inclined to contribute to democratic stability?

When is Democracy Over Inferred Preferences Valid?
Luke Thorburn, *King's College London*
Karen Ullrich, *Meta FAIR*
Max Nickel, *Meta FAIR*

If we replace votes with predicted votes, is it still a democracy? This paper is about the use of inferred preferences—preferences that are not explicitly stated but are instead extrapolated from what is already known about a voter—in democratic decision-making. First, the authors assess whether (and under what conditions) the use of inferred preferences is compatible with common notions of democracy and political representation. Second, they characterize what kinds of guarantees on the accuracy of such inferences are (a) desirable and (b) technically possible. They conclude with concrete recommendations for those designing and using tools for preference aggregation.

AI, Multiracial Democracy, Community Power, and Philanthropy
Spencer Overton, *George Washington University*

Demographic and technological shifts have placed the U.S. at a crossroads in defining its national identity. This essay explores how AI applications can entrench the past racial allocation of political power by facilitating bias, homogenization, cultural anxiety, racial polarization, and voter suppression. It also examines AI's potential to advance a more racially inclusive, pluralistic democracy by enhancing voting rights enforcement, language access, civic engagement, and voter mobilization. Finally, it details the crucial roles of community power-building groups, civil rights organizations, and philanthropy in ensuring that AI is leveraged to strengthen rather than weaken multiracial democracy.



PANEL 5 - FRIDAY

Reframing AI in the Digital Public Sphere

9:45 am - 10:55 am

Democracies depend on a vibrant public sphere. No social institution has been more shaped by narrow AI than the digital public sphere—by most accounts to its detriment. This panel asks if more capable general purpose AI systems can do better.

Panelists
Sacha Altay, *University of Zurich*
Joshua P. Darr, *Syracuse University*
Manon Revel, *Meta FAIR*
Ranjit Singh, *Data & Society*

Moderator
Nadine Farid Johnson, *Knight First Amendment Institute*

PAPER ABSTRACTS

Don't Panic (Yet): Assessing the Evidence and Discourse Around (Generative) AI and Elections

Felix Simon, *University of Oxford*
Sacha Altay, *University of Zurich*

This paper critically examines the impact of generative artificial intelligence on elections. Despite concerns about AI enabling more efficient voter targeting and persuasion, as well as more effective mis- and disinformation, the authors argue that AI's influence on election outcomes has been overestimated. They identify several factors behind this overestimation and analyze the skewed discourse on AI's role in elections, framing it as part of an ongoing "cycle of technology panics." They argue that focusing on AI may distract from more structural threats to elections and democracy, including voter disenfranchisement and attacks on election integrity, and call for a more nuanced approach that considers AI within broader sociopolitical contexts.

The Potential of Participatory Journalism for AI-Assisted Local News

Joshua P. Darr, *Syracuse University*

This paper proposes a civically useful way to incorporate AI into local journalism by prioritizing the concerns of community residents. Using data from a U.S. city to create local news AI chatbots, it demonstrates why the inputs to a local news AI matter. It warns against training local news AI on readily available data, such as newspaper archives and press releases, which may perpetuate traditional news values that do not always help people navigate their government or strengthen democracy. The paper further suggests how new approaches to journalism could improve civic outcomes, including transparency, participation, and public engagement.

Representative Ranking for Deliberation in the Public Sphere
Manon Revel, *Meta FAIR*
Smitha Milli, *Meta FAIR*
Tyler Lu, *Meta*
Jamelle Watson-Daniels, *Meta FAIR*
Max Nickel, *Meta FAIR*

Online comment sections have the potential to foster informal public deliberation. However, these settings often invite toxic or low-quality exchanges. Algorithmic ranking is increasingly deployed to facilitate higher-quality discussions, but these interventions may also inadvertently censor legitimate viewpoints, undermining another key aspect of deliberation: representation of diverse views. This paper seeks to remedy this problem by introducing guarantees of representation. The authors adopt a notion of justified representation (JR) from the social choice literature and incorporate a JR constraint into the comment ranking setting. They find that enforcing JR leads to greater inclusion of diverse viewpoints while still being compatible with optimizing for user engagement or other measures of conversational quality.

Experimental Publics: Democracy and the Role of Publics in GenAI Evaluation
Jacob Metcalf, *Data & Society*
Ranjit Singh, *Data & Society*
Borhane Blili-Hamelin, *AI Risk and Vulnerability Alliance*

This paper examines how emerging public experiments in red-teaming generative AI systems engender new formations of "experimental publics," shaped by competing visions of expertise, democratic oversight, and AI safety. Drawing on interviews with red-teaming practitioners, the authors analyze distinctions between instrumental and deliberative public feedback, situating them within the historical Dewey-Lippmann debate on democratic governance. By bridging AI evaluation with political epistemology, they explore how public involvement both challenges and reshapes the governance of generative AI systems, revealing its democratic potential as well as its contested boundaries.



PANEL 6

AI, Freedom of Expression, and Democratic Backsliding

11:10 am - 12:20 pm

As large language models become more and more integrated into the information environment, alignment techniques designed to limit AI companies' exposure to liability and public embarrassment are affecting the expressive interests of model users (including jailbreakers). At the same time, the community of research on AI risks is facing unprecedented political intervention into the content and methods of its research. All this is happening at a moment of rising authoritarianism and democratic backsliding. This panel asks how to square the values of safe AI with core principles of freedom of expression—preserving expressive interests against both private and state control.

Panelists

Becca Branum, *Center for Democracy and Technology*
Aziz Huq, *University of Chicago*
Xiangnong (George) Wang, *Knight First Amendment Institute*

Moderator

Katy Glenn Bass, *Knight First Amendment Institute*



KEYNOTE ADDRESS

Erica Chenoweth

1:00 pm - 2:00 pm

Erica Chenoweth is the academic dean for faculty development and the Frank Stanton Professor of the First Amendment at Harvard Kennedy School, faculty dean at Pforzheimer House at Harvard College, and a Susan S. and Kenneth L. Wallach Professor at Harvard's Radcliffe Institute. Chenoweth studies political violence and its alternatives and has authored or edited numerous books and dozens of articles on mass movements, nonviolent resistance, terrorism, political violence, revolutions, and state repression. At Harvard, Chenoweth directs the Nonviolent Action Lab, an innovation hub that uses social science tools and evidence to support movement-led political transformation. Chenoweth has taught at the University of Denver and Wesleyan University. They hold a Ph.D. and an M.A. in political science from the University of Colorado and a B.A. in political science and German from the University of Dayton.



PANEL 7

Transitions to Transformative AI

2:15 pm - 3:25 pm

Even if AI capabilities stalled out at their present level, the societal ramifications would be massive. But if progress continues at roughly the same pace, we may be dealing with even more profoundly transformative AI systems within the decade. This panel seeks to get on the front foot—to characterize what 'transformative AI' would mean, how it would differ from what has been realized so far, and how to develop societal resilience to the potential for further radical social change.

Panelists

Justin Bullock, *University of Washington*
Rumman Chowdhury, *Humane Intelligence*
Séb Krier, *Google DeepMind*
Sam Manning, *Center for the Governance of AI*

Moderator

Seth Lazar, *Australian National University and the Knight First Amendment Institute*

SPEAKERS

Sacha Altay
Sacha Altay is an experimental psychologist working on misinformation, trust, and social media in the Digital Democracy Lab at the University of Zurich. Through his research, he tries to understand why, despite people's socio-cognitive abilities to resist misinformation, some false beliefs are widespread, and he tests novel communication techniques to inform people efficiently and correct common misperceptions about vaccines, GMOs, and nuclear energy. Altay completed his Ph.D. in cognitive science at the École Normale Supérieure in Paris and was previously a postdoctoral research fellow at the Reuters Institute for the Study of Journalism at the University of Oxford.

Becca Branum
Becca Branum is the deputy director of the Free Expression Project at the Center for Democracy and Technology (CDT). Prior to joining CDT, Branum served as a senior counsel in the U.S. Senate. She began her career at the Consortium on Law and Values in Health, Environment & the Life Sciences at the University of Minnesota, where she focused on the legal and ethical implications of emerging technologies. She also previously served as a law clerk for the Senate Judiciary Subcommittee on Privacy, Technology, and the Law. Branum is a graduate of Boston University and the University of Minnesota Law School.

Justin Bullock
Justin Bullock is an associate professor affiliate of governance at the Evans School of Governance at the University of Washington. He is a scholar and expert in public policy, governance, public administration, technology, artificial intelligence, management, and organizational theory. His recent work examines the impact of technology on governance with a specific focus on the new forms of governance created by artificial intelligence and the growth of intelligent machines. He also explores the myriad ways in which new forms of technology, including the rise of the Metaverse, present new opportunities for education and spreading knowledge. Previously, Bullock was a tenured professor at Texas A&M University in the Busch School of Government and Public Service.

Rumman Chowdhury
Rumman Chowdhury's passion lies at the intersection of artificial intelligence and humanity. She is a pioneer in the field of applied algorithmic ethics, creating cutting-edge sociotechnical solutions for ethical, explainable, and transparent AI. She is the CEO and co-founder of Humane Intelligence and the first person to be appointed as the U.S. Science Envoy for Artificial Intelligence. Chowdhury

was the director of META (ML Ethics, Transparency, and Accountability) team at Twitter, leading a team of applied researchers and engineers to identify and mitigate algorithmic harms on the platform. Prior to Twitter, she was the CEO and founder of Parity, an enterprise algorithmic audit platform company, and served as the global lead for responsible AI at Accenture Applied Intelligence. Chowdhury holds two undergraduate degrees from MIT, a master's degree in quantitative methods of the social sciences from Columbia University, and a doctorate in political science from the University of California, San Diego.

Beba Cibralic

Beba Cibralic is a researcher at RAND, focusing on AI governance. Previously, she was a visiting scholar at the Leverhulme Centre for the Future of Intelligence at the University of Cambridge. She is the co-author of *Machine Agency* (2025). Cibralic has worked on AI ethics in industry, as a responsible AI lead at JPMorgan Chase & Co., and contributed to AI policy through projects with the Center for Democracy and Technology and the Center for Security and Emerging Technology. She received a Ph.D. in philosophy from Georgetown University, an M.A. in China studies from Peking University, where she studied as a Yenching Scholar, and an undergraduate degree in philosophy and political science from Wellesley College.

Joshua P. Darr

Joshua P. Darr is an associate professor at the Newhouse School of Public Communications and a senior researcher at the Institute for Democracy, Journalism & Citizenship at Syracuse University, teaching classes in political communication. His research focuses on local news, partisan polarization, and campaign strategy. Darr is the co-author of two books—*Home Style Opinion: How Local Newspapers Can Slow Polarization* (2021) and *Storefront Campaigning* (2024). He was the recipient of the Andrew Carnegie Fellowship in 2022 for his project investigating how newspapers can earn back trust and reduce political polarization. Darr earned a Ph.D. in political science from the University of Pennsylvania and a B.A. in political science from Boston College.

SPEAKERS

George Deodatis

George Deodatis is vice dean for Research at Columbia University's School of Engineering and Applied Science, Santiago and Robertina Calatrava Family Professor of Civil Engineering and Engineering Mechanics, and professor of Earth and Environmental Engineering. He uses probabilistic methods and uncertainty quantification for the study of civil infrastructure systems subjected to natural and technological hazards. He conducts research to determine the safety and reliability of structures and to perform risk assessment and risk management of structural systems. He also studies the effects of climate change and extreme weather on the civil infrastructure with the goal of establishing adaptation and mitigation strategies.

Nadine Farid Johnson

Nadine Farid Johnson is the inaugural policy director of the Knight First Amendment Institute. She is responsible for the Institute's policy and advocacy efforts, leading engagement with U.S. government and other officials to advance the Institute's policy objectives. A former U.S. diplomat and professor of law and political science, Farid Johnson is a frequent media contributor, with commentary appearing in The New York Times, The Wall Street Journal, The Atlantic, World Politics Review, Al Jazeera, The Hill, The Daily Beast, NPR, and other national and international publications. Farid Johnson comes to the Knight Institute most recently from PEN America, where she served as the managing director of that organization's Washington office and Free Expression programs. She is a graduate of DePauw University and Tulane Law School and studied at the U.S. Naval War College.

Kevin Feng

Kevin Feng is a Ph.D. student in the Department of Human Centered Design and Engineering at the University of Washington. His research interests lie at the intersection of social computing and human-centered AI. Specifically, his work focuses on designing new sociotechnical methods and interactive systems to advance AI usability, safety, and governance. His work is supported by an OpenAI Democratic Inputs to AI grant (2023) and a UW Herbold Fellowship (2022). Feng has held internships and part-time research roles at the Allen Institute for AI (Ai2) and at Microsoft Research. He received a B.S.E. in computer science from Princeton University, where he conducted research at the Center for Information Technology Policy.

Katy Glenn Bass
Katy Glenn Bass is the research director of the Knight First Amendment Institute. She is responsible for conceptualizing and executing the Institute's research initiatives, including the production of scholarship and research materials, the organization of conferences and symposia, and the Institute's Visiting Research Scholars program. Glenn Bass previously worked at PEN America, where she supervised the production of reports analyzing free expression issues, including the lack of protections for national security whistleblowers, attacks on journalists covering the 2014 protests in Ferguson, Missouri, and the impact of mass surveillance on online speech. She is a graduate of Princeton University and Harvard Law School.

Gillian Hadfield
Gillian Hadfield is a professor of computer science at the Whiting School of Engineering at Johns Hopkins University and has a joint appointment in the School of Government and Policy. She holds a CIFAR AI Chair at the Vector Institute for Artificial Intelligence and is a Schmidt Sciences AI2050 Senior Fellow. Originally trained as an economist and a legal scholar, Hadfield now focuses her research on collaborations with machine learning researchers to build systems that understand and respond to human norms and on innovative design for legal and regulatory systems for artificial intelligence and other complex global technologies.

Hahrie Han
Hahrie Han is the inaugural director of the SNF Agora Institute, the Stavros Niarchos Foundation Professor of Political Science, and the faculty director of the P3 Research Lab at Johns Hopkins University. She is an award-winning author of five books and numerous scholarly articles. Her latest book about faith and race in America with a focus on evangelical megachurches was named to The New York Times list of 100 Notable Books of the Year in 2024 and The New Yorker's list of Recommended Books for 2024. Her writing has appeared in scholarly and public outlets, including The New York Times, The Washington Post, and the Journal of the American Medical Association. She is an elected member of the American Academy of Arts and Sciences, was named a 2022 Social Innovation Thought Leader of the Year by the World Economic Forum's Schwab Foundation, and delivered the Tanner Lectures at Harvard University in 2024.

SPEAKERS

Hoda Heidari

Hoda Heidari is the K&L Gates Career Development Assistant Professor in Ethics and Computational Technologies at Carnegie Mellon University with joint appointments in the machine learning department and the Institute for Software, Systems, and Society. She is also affiliated with the Human-Computer Interaction Institute and the Heinz College of Information Systems and Public Policy at Carnegie Mellon University and co-leads the university-wide Responsible AI Initiative and the K&L Gates Initiative for Ethics and Computational Technologies. Heidari is interested in the ethical and societal aspects of artificial intelligence and machine learning. In particular, her research has addressed issues of fairness and accountability. She is part of the AI Institute for Societal Decision Making, one of the National Artificial Intelligence Research Institutes funded by the National Science Foundation.

Peter Henderson

Peter Henderson is an assistant professor at Princeton University with joint appointments in the department of computer science and the School of Public and International Affairs. His research lies at the intersection of artificial intelligence and law, focusing on developing safe, capable, and efficient AI systems that benefit society, as well as leveraging law and evidence-based policy to steer toward positive outcomes for AI. He holds a J.D. from Stanford Law School and a Ph.D. in computer science from Stanford University. His research has been used by companies and government agencies across the world and has been featured in The New York Times, The Wall Street Journal, Science, Bloomberg, and more.

Aziz Huq

Aziz Huq is the Frank and Bernice J. Professor at the University of Chicago Law School and an associate professor of sociology. His books include *How to Save a Constitutional Democracy* (2018), co-authored with Tom Ginsburg; *The Collapse of Constitutional Remedies* (2021); and *The Rule of Law: A Very Short Introduction* (2024). His scholarship has been published in leading law, social science, political science, and criminology journals. He also writes for Politico, the London Review of Books, and many other general interest publications. Previously, Huq represented civil liberties claimants with the Brennan Center for Justice and worked for the International Crisis Group in Afghanistan, Nepal, Pakistan, and Sri Lanka. He was also a law clerk for Judge Robert D. Sack (U.S. Court of Appeals for the Second Circuit) and Supreme Court Justice Ruth Bader Ginsburg.

Lujain Ibrahim

Lujain Ibrahim is a Ph.D. candidate in social data science at the University of Oxford. Her research investigates sociotechnical approaches to evaluating and governing AI systems, their societal impact, and their relational harms. Formerly, she was a student researcher at Google DeepMind, a Schwarzman Scholar, and a fellow at the Centre for Governance of AI and the Digital Asia Hub. She holds an M.A. in global affairs with concentrations in technology and policy from Tsinghua University and a B.S. in computer engineering from New York University, Abu Dhabi.

Jameel Jaffer

Jameel Jaffer is the executive director of the Knight First Amendment Institute. Under his leadership, the Institute has filed precedent-setting litigation, undertaken major interdisciplinary research initiatives, and become an influential voice in debates about the freedoms of speech and the press in the digital age. He is an executive editor of Just Security, a national security blog, and his most recent book, *The Drone Memos*, was one of The Guardian's "Best Books of 2016." Until August 2016, Jaffer served as deputy legal director at the ACLU, where he oversaw the organization's work on free speech, privacy, technology, national security, and international human rights. He litigated many significant post-9/11 cases involving human rights and national security, including cases relating to surveillance, secrecy, censorship, interrogation, detention, and extrajudicial killing. He was a distinguished fellow at the University of Toronto's Munk School from 2016 to 2021 and a fellow at the Open Society Foundations in 2013.

Atoosa Kasirzadeh

Atoosa Kasirzadeh is a philosopher and AI researcher focusing on ethics and governance of AI and computing. She is an assistant professor at Carnegie Mellon University with joint affiliations in the philosophy and software and societal systems departments, a Schmidt Sciences AI2050 Early Career Fellow, a research affiliate with the Centre for the Governance of AI, a steering committee member for the ACM FAccT conference, and a council member of the World Economic Forum's Global Future Council on Artificial General Intelligence. Previously, Kasirzadeh was a visiting faculty at Google Research, a Chancellor's Fellow and research lead at the University of Edinburgh's Centre for Technomoral Futures, a group research lead at the Alan Turing Institute, a DCMS/UKRI senior policy fellow, and a Governance of AI Fellow at the University of Oxford. Kasirzadeh holds doctorates in philosophy of science and technology from the University of Toronto and mathematics from the École Polytechnique de Montréal, and an M.Sc. and a B.Sc. in systems engineering.

SPEAKERS

Séb Krier

Sébastien (Séb) Krier leads policy development and strategy at Google DeepMind. He is an artificial intelligence policy expert, adviser, and attorney working to improve the regulatory and legal frameworks governing AI and studying the impacts of emerging technologies on democracy and human rights. He previously served in the U.K. government's Office for Artificial Intelligence as the head of regulation where he led the world's first comprehensive review of AI in the public sector and designed policies to address novel issues such as the oversight of automated decision-making systems. He also represented the United Kingdom at various multilateral fora such as the European Commission, the World Economic Forum, the OECD, and the Digital Nations. Krier studied law at King's College London and obtained an MPA focusing on technology and social impact from the University College London.

Seth Lazar

Seth Lazar is a professor of philosophy at the Australian National University, an Australian Research Council Future Fellow, and a Distinguished Research Fellow of the University of Oxford Institute for Ethics in AI. He has worked on the ethics of war, self-defense, and risk and now leads the Machine Intelligence and Normative Theory Lab, where he directs research projects on normative philosophy of computing. He was the general co-chair for the Association for Computing Machinery (ACM) Fairness, Accountability, and Transparency conference in 2022 and the program co-chair for the ACM/Association for the Advancement of Artificial Intelligence's AI, Ethics, and Society conference in 2021. He is one of the authors of a study by the U.S. National Academies of Science, Engineering, and Medicine on the ethics and governance of responsible computing research. He gave the 2022 Mala and Solomon Kamm lecture in ethics at Harvard University and the 2023 Tanner Lectures on AI and human values at Stanford University. For the 2024-2025 academic year, Lazar is a senior AI advisor at the Knight First Amendment Institute.

Sydney Levine

Sydney Levine is a research scientist at Google DeepMind focusing on AI safety. She uses the tools and methods of cognitive science to help engineer AI systems to be better aligned with the plurality of human moral values. Levine completed her Ph.D. in the Department of Psychology and the Center for Cognitive Science at Rutgers University and then held a joint appointment as a postdoctoral researcher in the Department of Psychology at Harvard University and the MIT Department of Brain and Cognitive Sciences. She will start a position as an assistant professor

in the New York University Department of Psychology in the spring of 2026, where her lab will work on issues at the intersection of moral cognition and AI safety.

Shayne Longpre
Shayne Longpre is a Ph.D. candidate at the Massachusetts Institute of Technology. His research focuses on the intersection of AI and policy: responsibly training, evaluating, and governing general-purpose AI systems. Longpre leads the Data Provenance Initiative, led the Open Letter on A Safe Harbor for Independent AI Evaluation & Red Teaming, and has contributed to training models like Bloom, Aya, and Flan-T5/PaLM. He is thankful for the recognition his research has received: Best Paper Awards from ACL 2024, NAACL 2024, as well as coverage by The New York Times, The Washington Post, The Atlantic, 404 Media, Vox, and MIT Tech Review.

Sam Manning
Sam Manning is a senior research fellow at the Centre for the Governance of AI. His work focuses on measuring the economic impacts of frontier AI systems and designing policy options to help ensure that advanced AI can foster broadly shared economic prosperity. He previously conducted research at OpenAI and worked on a randomized controlled trial of a guaranteed income program in the U.S. Manning has a M.Sc. in International and Development Economics from the University of San Francisco.

Jakob Mökander
Jakob Mökander is the director of Science & Technology Policy at the Tony Blair Institute for Global Change (TBI) and a research fellow at Yale University's Digital Ethics Center. Mökander has published extensively in the fields of tech policy, AI governance, and AI auditing. His work has been featured in journals like Nature Machine Intelligence, Science and Engineering Ethics, and Minds and Machines. Mökander holds a Ph.D. from the Oxford Internet Institute. Prior to joining TBI, he was a visiting scholar at Princeton University's Center for Information Technology Policy, and a research fellow at the Centre for the Governance of AI. Outside academia, Mökander has worked as a management consultant within McKinsey & Company's Risk and Resilience practice, as well as a trade and investment advisor at the Embassy of Sweden in New Delhi, India.

SPEAKERS

Deirdre K. Mulligan

Deirdre K. Mulligan is a professor in the School of Information at the University of California, Berkeley, a faculty director of the Berkeley Center for Law & Technology, a co-organizer of the Algorithmic Fairness & Opacity Group, an affiliated faculty on the Hewlett-funded Center for Long-Term Cybersecurity and a faculty advisor for its AI Policy Hub, and a faculty advisor to the CITRIS Policy Lab. Mulligan's research explores legal and technical means of protecting values such as privacy, freedom of expression, and fairness in sociotechnical systems. Mulligan served as the principal deputy U.S. chief technology officer at the White House Office of Science and Technology Policy and the director of the National Artificial Intelligence Initiative Office in the Biden-Harris administration. She previously served as staff counsel at the Center for Democracy and Technology. She received a J.D. from Georgetown Law School and a B.A. from Smith College.

Arvind Narayanan

Arvind Narayanan is a professor of computer science at Princeton University and the director of the Center for Information Technology Policy. He co-authored a textbook on fairness and machine learning and a book titled *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference* (2024). He led the Princeton Web Transparency and Accountability Project to uncover how companies collect and use personal information. His work was among the first to show how machine learning reflects cultural stereotypes, and his doctoral research showed the fundamental limits of de-identification. Narayanan is a recipient of the Presidential Early Career Award for Scientists and Engineers, twice a recipient of the Privacy Enhancing Technologies Award, and thrice a recipient of the Privacy Papers for Policy Makers Award. Narayanan was the Knight First Amendment Institute's 2022-2023 visiting senior research scientist.

Alondra Nelson

Alondra Nelson is the Harold F. Linder Professor at the Institute for Advanced Study and a Distinguished Senior Fellow at the Center for American Progress. As a former deputy assistant to President Joe Biden, she served as the principal deputy director for science and society and the acting director of the White House Office of Science and Technology Policy from 2021 to 2023. Nelson's work drove the Biden-Harris administration strategy to create science and technology policy that expands economic opportunity, protects civil rights, enhances security, advances equity, and ensures that innovation works for, not against, our democratic values. During her tenure, Nelson also led a team in writing the landmark

Blueprint for an AI Bill of Rights. An acclaimed social scientist, Nelson writes and lectures widely on the intersections of science, technology, medicine, and social inequality. She is the author of several books, including *The Social Life of DNA*.

Spencer Overton
Spencer Overton is the Patricia Roberts Harris Research Professor of Law at George Washington University. He is the author of the book *Stealing Democracy: The New Politics of Voter Suppression*, the law review article "Overcoming Racial Harms to Democracy from Artificial Intelligence," and several other publications on democracy and race. At the George Washington University Law School, he directs the Multiracial Democracy Project, which is currently working on research projects on the implications of artificial intelligence and alternative election systems for truly representative democracy in the United States. Overton held several senior leadership roles during the Obama campaign, transition, and administration. He currently serves on the board of the Leadership Conference Education Fund, the education and research arm of the Leadership Conference on Civil and Human Rights. He graduated with honors from both Hampton University and Harvard Law School.

Manon Revel
Manon Revel is a postdoctoral AI researcher at Meta FAIR and a research affiliate at the Berkman Klein Center for Internet & Society at Harvard University. Her research, spanning social choice theory and political philosophy, provides mathematical frameworks to analyze and enhance democratic processes online and offline. Revel graduated from MIT with a Ph.D. in social and engineering systems and statistics and an M.Sc. in technology and policy. She also holds an M.Sc. in applied mathematics from École Centrale Paris.

Ranjit Singh
Ranjit Singh is a senior researcher with the AI on the Ground team at the Data & Society (D&S) research institute. His work explores the ordinary ethics of exercising human agency in navigating AI-driven systems. He develops robust methods to evaluate and regulate the impact of algorithms on everyday life and helps guide research ethics at D&S, fostering equity in collaborative research both internally and with external partners. His research sits at the intersection of AI infrastructures, majority-world scholarship, and public policy, using interview-based qualitative sociology and multi-sited ethnography.

SPEAKERS

Daniel Susskind
Daniel Susskind is a research professor in economics at King's College London, a senior research associate at the Institute for Ethics in AI at the University of Oxford, and an associate member of the economics department at the University of Oxford. His work explores the impact of technology, particularly artificial intelligence, on work and society. Susskind is the author of *Growth: A Reckoning* (2024) and *A World Without Work* (2020), described by The New York Times as "required reading for any potential presidential candidate thinking about the economy of the future." He is also the co-author of the bestselling book *The Future of the Professions* (2015). Previously, he held various roles in the British government, including in the Prime Minister's Strategy Unit, the Policy Unit at 10 Downing Street, and the Cabinet Office. He was a Kennedy Scholar at Harvard University.

MH Tessler
Michael Henry (MH) Tessler is a senior research scientist at Google DeepMind based in London, working on the safety and alignment of large language models (LLMs). He is deeply interested in the potential for LLMs to support and scale human deliberation in the service of strengthening democracy. Before becoming an AI researcher, Tessler studied human language use and language understanding as a postdoctoral researcher at MIT and as a Ph.D. student in the Department of Psychology at Stanford University. His research has been featured in journals such as Science, Nature Human Behaviour, and Psychological Review.

Luke Thorburn
Luke Thorburn is a Ph.D. student at King's College London and a research fellow at the AI & Democracy Foundation. He works on topics at the intersection of algorithms and conflict, including systemic risks of social media algorithms, democratic infrastructure for AI governance, and the use of algorithms to support peace processes in conflict regions.

Helen Toner
Helen Toner is the director of strategy and foundational research grants at Georgetown University's Center for Security and Emerging Technology (CSET). She previously worked as a senior research analyst at Open Philanthropy, where she advised policymakers and grantmakers on AI policy and strategy. Between working at Open Philanthropy and joining CSET, Toner lived in Beijing, studying the Chinese AI ecosystem as a research affiliate of the University of Oxford's Centre for the Governance of AI. Toner has written for Foreign Affairs and other

outlets on the national security implications of AI and machine learning for China and the United States, and has testified before the U.S.-China Economic and Security Review Commission. Toner holds an M.A. in security studies from Georgetown University, as well as a B.Sc. in chemical engineering and a diploma in languages from the University of Melbourne.

Xiangnong (George) Wang

Xiangnong (George) Wang is a staff attorney at the Knight First Amendment Institute. He first joined the Institute as a legal fellow from 2020 to 2021 before serving as a law clerk to the Hon. William J. Kayatta Jr. of the U.S. Court of Appeals for the First Circuit. Wang received a B.A. in public policy from Stanford University and a J.D. from Yale Law School. During law school, he was a student director of the Media Freedom & Information Access Clinic, co-chair of the Asian Pacific American Law Students Association, and a Coker Fellow. Wang also worked on research related to improving diversity and equity in the legal profession as a member of the Portrait of Asian Americans in the Law project. In a previous career, he worked as an engineer. He has been published in the Wisconsin Law Review, the Journal of Legal Education, and Just Security.



The Knight First Amendment Institute at Columbia University defends the freedoms of speech and the press in the digital age through strategic litigation, research, policy advocacy, and public education. It promotes a system of free expression that is open and inclusive, that broadens and elevates public discourse, and that fosters creativity, accountability, and effective self-government.

The Knight Institute is grateful for the support of Arnold Ventures; the Bernard and Anne Spitzer Charitable Trust; the Carnegie Corporation of New York; the Charles Koch Foundation; Columbia University; David Bohnett Foundation; Abby and Rick Elbaum; the Ford Foundation; the Ford Foundation Dignity and Justice Fund at New Venture Fund; Hearst Foundations; the Helen Gurley Brown Foundation; the Internet and Platform Researcher First Amendment Support Fund at The Miami Foundation; the John D. and Catherine T. MacArthur Foundation; Abby and Jonathan Moses; the John S. and James L. Knight Foundation; Netgain Partnership, a project of Media Democracy Fund; The Omidyar Group; Open Society Foundations; the Pritzker Pucker Family Foundation; the William and Flora Hewlett Foundation; and the Zegar Family Foundation.

knightcolumbia.org

The Knight Institute would also like to thank Columbia Engineering for their support and partnership.

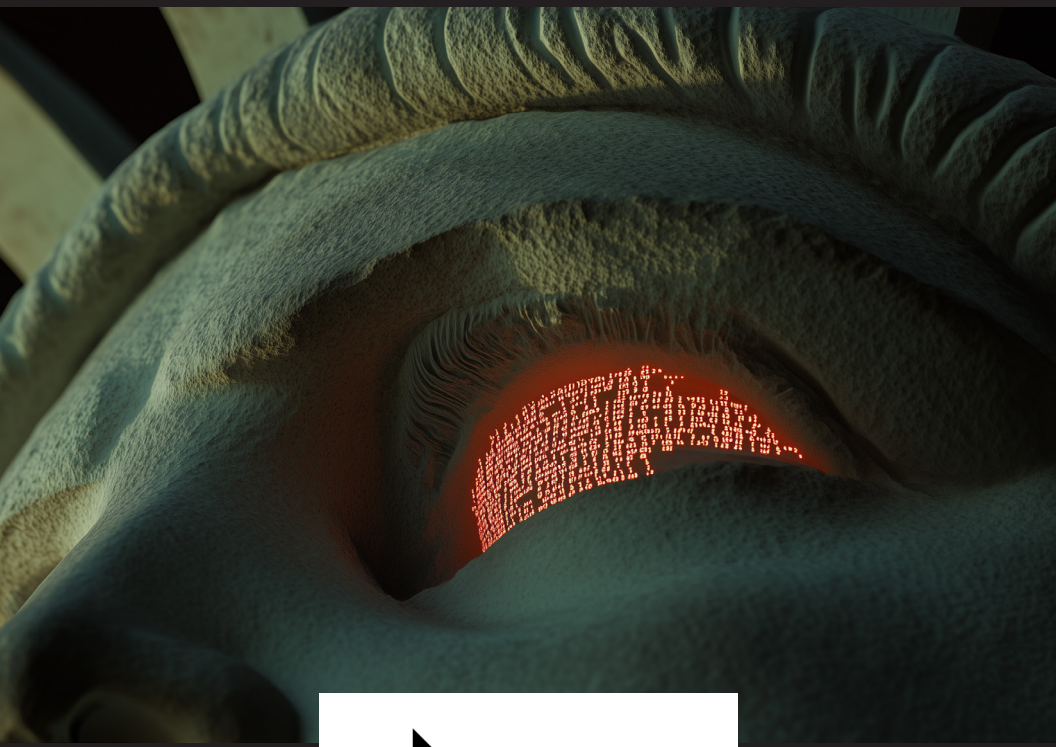


Illustrations: Sébastien A. Krier using Midjourney 6.1 | Design: Astrid Da Silva

Please use the QR code below to ask questions during the presentations.



If you plan on posting about this event on social media, please include the hashtag **#AIDemocraticFreedoms**



KNIGHT
FIRST AMENDMENT
INSTITUTE at
COLUMBIA UNIVERSITY