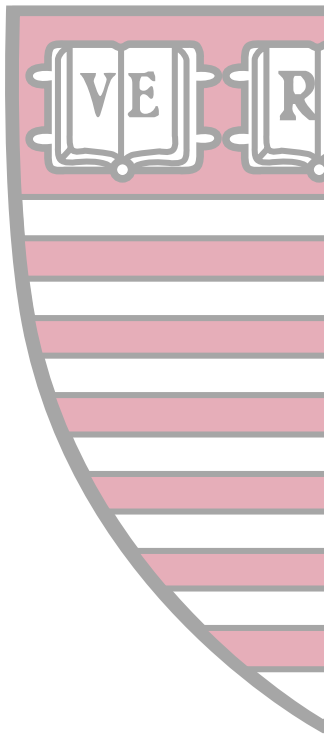
OCCASIONAL PAPERS SERIES



A Roadmap for Governing AI: Technology Governance and Power Sharing Liberalism

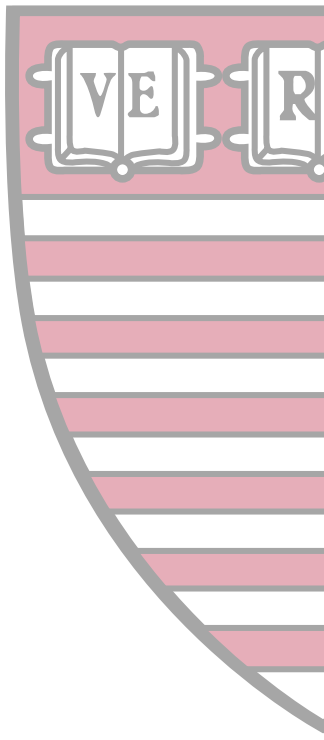
Danielle Allen, Sarah Hubbard, Woojin Lim, Allison Stanger, Shlomit Wagman, and Kinney Zalesne

January 2024
Preprint version

A product of the Allen Lab for Democracy Renovation

HARVARD Kennedy School
ASH CENTER
for Democratic Governance
and Innovation

OCCASIONAL PAPERS SERIES



A Roadmap for Governing AI: Technology Governance and Power Sharing Liberalism

Danielle Allen, Sarah Hubbard, Woojin Lim, Allison Stanger, Shlomit Wagman, and Kinney Zalesne

January 2024
Preprint version

A product of the Allen Lab for Democracy Renovation

Abstract

This paper aims to provide a roadmap to AI governance. In contrast to the reigning paradigms, we argue that AI governance should not be merely a reactive, punitive, status-quo-defending enterprise, but rather the expression of an expansive, proactive vision for technology—to advance human flourishing. Advancing human flourishing in turn requires democratic/political stability and economic empowerment. Our overarching point is that answering questions of how we should govern this emerging technology is a chance not merely to categorize and manage narrow risk but also to construe the risks and opportunities much more broadly, and to make correspondingly large investments in public goods, personnel, and democracy itself. To lay out this vision, we take four steps. First, we define some central concepts in the field, disambiguating between forms of technological harms and risks. Second, we review normative frameworks governing emerging technology that are currently in use around the globe. Third, we outline an alternative normative framework based in power-sharing liberalism. Fourth, we walk through a series of governance tasks that ought to be accomplished by any policy framework guided by our model of power-sharing liberalism. We follow these with proposals for implementation vehicles.

Keywords: Artificial Intelligence, governance, liberalism, pluralism, political economy, tech ethics

About the Authors

Danielle Allen is the James Bryant Conant University Professor at Harvard University and Director of the Allen Lab for Democracy Renovation at Harvard Kennedy School's Ash Center for Democratic Governance and Innovation. She is a professor of political philosophy, ethics, and public policy.

Sarah Hubbard is a product leader and technologist. She is currently a Senior Fellow with the Allen Lab for Democracy Renovation at the Ash Center for Democratic Governance and Innovation.

Woojin Lim is a Harvard College student in Philosophy and Government and a member of the Allen Lab for Democracy Renovation.

Allison Stanger is the Co-Director of the GETTING-Plurality Research Network at the Allen Lab for Democracy Renovation and the Russell Leng '60 Professor of International Politics and Economics at Middlebury College.

Shlomit Wagman is a Faculty Associate at the Berkman-Klein Center at Harvard Law School, a Research Fellow at Havard Kennedy School's Mossavar-Rahmani Center for Business and Government, and a member of the GETTING-Plurality Research Network at the Allen Lab for Democracy Renovation.

Kinney Zalesne is a member of the GETTING-Plurality Research Network at the Allen Lab for Democracy Renovation.

About the Ash Center

The Mission of the Roy and Lila Ash Center for Democratic Governance and Innovation at is to develop ideas and foster practices for equal and inclusive, multi-racial and multi-ethnic democracy and self-government.

This paper is one in a series published by the Ash Center for Democratic Governance and Innovation at Harvard University's John F. Kennedy School of Government. The views expressed in the Ash Center Policy Briefs Series are those of the author(s) and do not necessarily reflect those of the John F. Kennedy School of Government or of Harvard University. The papers in this series are intended to elicit feedback and to encourage debate on important public policy challenges.

About the Allen Lab for Democracy Renovation

The Allen Lab for Democracy Renovation seeks to address the threats to American and global democracies with research and field-building to support a positive vision of what Danielle Allen has called "power-sharing liberalism." This vision is grounded in the belief that just societies require robust political equality, fully inclusive institutions, and broader avenues for participation and connectedness, all of which will rest in turn on the material and social bases for human flourishing. To those ends, our multidisciplinary community of scholars, practitioners, and partner organizations work together to shepherd concepts and reforms into practice—to translate research into impact. From community-led initiatives to national-level policies and structural reforms, the Allen Lab seeks to renovate American democracy.

A PUBLICATION OF THE

Ash Center for Democratic Governance and Innovation
Harvard Kennedy School
79 John F. Kennedy Street
Cambridge, MA 02138

617-495-0557
ash.harvard.edu

Acknowledgements

We are grateful for the comments and feedback offered on earlier drafts of this paper by members of the GETTING-Plurality Research Network. In particular, we are especially grateful for feedback from and discussions with Alex Pascal, Nick Pyati, and Tantum Collins. We are also grateful for Eli Frankel's extraordinary research assistance.

Declarations

Ethics approval and consent to participate: No human subjects in research. Ethics approval not applicable.

Consent for publication: All authors consent and all materials used with permissions from authors and rights-holders.

Availability of data and material: All data and source material used with proper permissions and publicly available. In particular, this paper adapts or re-publishes some material from other papers from our research network, including "Putting Flourishing First: Applying Democratic Values to Technology," by Kinney Zalesne and Nick Pyati (2023), "Ethics of Decentralized Social Technologies" by Allen et al. (2023), and elements of policy memos the network submitted to the White House OSTP, all with permission from the authors.

Competing interests: Not applicable.

Funding: This research originated under the aegis of The GETTING Plurality Network at the Allen Lab, which is grateful for the support of the Omidyar Network Fund, Inc., the William and Flora Hewlett Foundation, EY, the Microsoft Corporation, and the Harvard Ash Center for Democratic Governance and Innovation.

Authors' contributions: All listed authors contributed evenly to this manuscript.

Contents

Section I. Introduction

Technological breakthroughs throughout human history have brought both opportunities and harms. From fire to gunpowder to nuclear power, each has made it possible for human beings to overcome barriers to flourishing, population growth, and sustenance. Each has also introduced terrible harms—from arson to increasingly destructive warfare and peacetime violence to the nuclear weapons. Social trajectories have been dramatically transformed by technological innovations. Before the cotton gin, some American political leaders believed that enslavement was a dying economic form; the cotton gin was partially responsible for reviving it (Beckert, 2014).

The historian Ian Morris makes the case that major social transformations considered over long arcs of history are best understood as stemming from changes in "geography": when something occurs to change how human beings experience time and space, many other human social structures will change around that (Morris, 2022). Social media has closed geographic distance, permitting people with shared views–even extreme ones—to find each other over great distances. This risks destabilizing institutions of democratic representation built on 18th century meanings of geography. With the recent advances in artificial intelligence, we may not yet be on the threshold of the emergence of intelligent silicone, but we have certainly now witnessed the arrival of a technology built off our own cultural output, which we do not understand well, which has some alien features, and which may change how we interact across time and space.

That novel technology raises important questions about our longstanding copyright system, about the reliability and security of our information ecosystems, about the concentration of power, and about the proper role of technology in self-governed democratic societies, to name just a few of the emerging issues. Already, artificial intelligence technologies built on machine learning have supercharged forms of discrimination and prejudice, contributed to entrenched political polarization, and challenged the credibility of public communication channels (Persily and Tucker, 2020; Guess and Lyons, 2020). While it is hardly controversial to bemoan the private control of contemporary technological progress—control that currently lies in the hands of a handful of Silicon Valley executives and investors—the issue of governing these novel technologies has remained quite divisive.

This paper aims to provide a roadmap to AI governance. In contrast to the reigning paradigms, we argue that AI governance should not be merely a reactive, punitive, status-quo-defending enterprise, but rather the expression of an expansive, proactive vision for technology—namely, to advance human flourishing. Advancing human flourishing in turn requires democratic and political stability and economic empowerment. Our overarching point is that answering questions of how we should govern this emerging technology is a chance not merely to categorize and manage narrow risk but also to construe the risks and opportunities much more broadly, and to make correspondingly large investments in public goods, personnel, and democracy itself.

To lay out this expansive vision, we take four steps. (1) We define some central concepts in the field, disambiguating between various forms of technological harms and risks. (2) We review normative frameworks governing emerging technology that are currently in use around the globe. (3) We outline our proposed normative framework to guide governance proposals, drawing from contemporary political philosophy, and specifically, Danielle Allen's argument for power-sharing liberalism in *Justice by Means of Democracy* (2023). (4) We outline a governance framework that would delegate responsibility for different modalities of harm and opportunity to various public institutions and democratic mechanisms.

For that fourth and final step, we walk through six governance tasks that must be accomplished by any successful policy framework for governing artificial intelligence: mitigating harm from both the production and use of new technologies; blocking bad actors; equipping ourselves to see and maintain human mastery over emergent capabilities; identifying opportunities and steering toward public goods; building a human capital strategy; and reinforcing and strengthening democratic steering capacity. We follow this task review with proposals for implementation vehicles.

Importantly, as we outline this governance framework, we are also addressing the active debate as to whether AI governance should be framed by thinking about already present near-term harms such as bias or about potential, existentially threatening risks. We argue that this is a false dichotomy, not a genuine philosophical question, and can be resolved through choices about organizational structure.

Throughout our paper, in order to make our general approach to governing AI concrete, we reflect on specific proposals that are emerging in the U.S. context (as in the recent U.S. Presidential Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence and the EU's AI Act) (EO 14110, 2023; EU Commission, 2023). We also make further specific proposals that suit the U.S. context. Nonetheless, our governance approach is not U.S. specific and could be translated into other structures of national governance. Because governing AI will ultimately require that vocabulary and core concepts be shared at a global level, we hope our proposal here can inform the development of that shared global vocabulary.

Section II. Key Concepts

This section reviews some of the basic concepts commonly used for thinking about governance of AI.[1]

Technical Concepts

Foundation Model: A foundation model refers to a large neural network which is notably adaptable and capable of a wide variety of tasks (e.g., OpenAI's GPT-3 and GPT-4) (Jones, 2023).

Open-Source: An open-source model allows anyone to publicly access, modify, and update the source code of the system. Examples include LLaMA and Falcon 40B (Bommasani et al., 2023). In contrast, some models are closed-source and private so that only certain companies or developers have access to the underlying source code like OpenAI's GPT-3 and DeepMind's AlphaZero (Bommasani et al., 2023). In addition, the middle ground of "source-available" or "partially open source" models involves arrangements where the source code can be viewed or in some cases modified and enhanced by anyone, while key components remain proprietary.

Alignment: Alignment generally refers to how a model is calibrated to a set of guiding rules or principles in order to steer towards intended output. This process may be done through fine-tuning and other methods.

Model Development Lifecycle: The engineering lifecycle for developing an AI model typically follows a series of activities centered around designing the model, training the model including iterations of tuning and testing, deploying the model, and monitoring the model using performance metrics which can cycle back into the next training phase (U.S. GSA, 2023; Shevlane et al., 2023).

Risk Frameworks[2]

Capabilities: This term encapsulates the functions that AI systems are able to perform. AI capabilities to date have included functions "such as classifying data (e.g., assigning labels to images), grouping data (e.g., identifying customer segments with similar purchasing behavior), making predictions, or choosing actions (e.g., steering an autonomous vehicle)" (Toner, 2023). In contrast to earlier AI systems,

foundation models can generate content at a new level of scale and are demonstrating novel capabilities such as complex reasoning and synthetic media. We are still quite far from a full understanding of the implications of this capability.

Use Cases: AI systems are put to use in specific "use cases": e.g., to support algorithmic prediction by insurance firms, judicial systems, or airline companies; to operate autonomous vehicles; to deliver answers to users seeking information through a chatbot. The question of whether an AI system delivers harm or benefit depends on the structure and operations of the use case in which the system is embedded.

Interaction Effects: Because AI systems are deployed in use cases that are themselves connected to other social and biological phenomena, effects of AI systems may emerge that flow not from the specific technical capabilities of the system but rather from how that capability interacts with other phenomena. Algorithmic prediction in social media might generate viewer addiction, but it is the impact of addiction, and the specific objects of addiction, that generate negative mental health outcomes in young people. The AI system itself may not have the capability to generate negative mental health outcomes; rather its capability to identify content most likely to retain attention *interacts* with elements of youth development and cultural context, together driving negative mental health outcomes.

Harm and Benefit to Individuals and Organizations vs Systemic Harm and Benefit: A basic rule of thumb can guide all thinking about new technology: We wish for technologies to avoid harming people and also to bring them benefit. More complicated, though, is that new technologies can have impacts on individuals or on human groups, including organizations, societies, or even the whole human race. Governing technology requires a framework for thinking about individual well-being and organizational well-being, but also collective well-being, so that potential harms and benefits can be governed on micro, meso, and macro scales. Coal-mining did harm to individual miners. As businesses, coal-mining firms have experienced their own distinctive evolutions over time, with changes in ownership models and structures, and viability. With regard to social impacts, coal-mining powered an economy that delivered growth and improved material security for billions of people; at the same time, the burning of fossil fuels has heated the climate and left all of humanity exposed to climate crisis.

Capabilities Risk and Interaction Risk: As we seek to govern AI, we must not only distinguish between harms and benefits at the level of the individual vs. the group (up to the scale of all of humanity), but also between the risks that flow from the capabilities themselves and those that flow from how the capabilities of an AI system interact with other biological and social systems. The use of a predictive AI system to make decisions about bail exhibits capability risk when the training of the system on data reflecting racial bias results in a set of decisions that reinforce that racial bias (Angwin et al., 2023). The use of such a system by judges who come to rely on it and increasingly give over swathes of their own decision-making power to algorithms can lead to undermining the perceived legitimacy of a court system and weaken judicial institutions. The former is a capabilities risk; the latter is an interaction risk.

Risk Framework and Risk Levels: Most AI governance efforts to date have focused on potential harms or risks. Like the European Union AI Act, frameworks typically focus on minimal, moderate, high, and unacceptable risk levels, assigning corresponding regulatory burdens to AI providers (EU Online, 2023). Regulators must provide substantive input and analysis to determine which use cases belong at each level. Unacceptable risks also often get denoted by three further terms: Catastrophic, Genocidal, and Existential Risk. Catastrophic Risk is the risk of an event or phenomenon that could overturn the whole structure and survival of a community. Genocidal Risk is the risk of an event or phenomenon

that could do the same for a specific ethnic, religious, or linguistic community. Existential Risk is the risk of an event that could do the same for all of humanity. Nuclear weapons, for instance, bring unacceptable risks at all three levels: catastrophic, genocidal, and existential.

Risk frameworks like the ones in the EU AI Act rely heavily on frameworks developed for governing human subjects research in the biomedical and behavioral sciences (Novelli et al., 2023).[3] Those frameworks have, historically, focused much more heavily on individual harms and benefits, rather than on harms and benefits to groups, societies, or humanity. Consequently, an important task for the development of AI governance strategies and tools is to expand existing risk frameworks to integrate attention to those scaled up challenges and opportunities. A diversity of vocabulary is used for scaled-up impacts that bring risks or opportunities: organizational risk, systemic risk and structural risk are three common terms (Zwetsloot and Dafoe, 2019; NIST 2023; Daníelsson et al., 2021).[4] Under this framing fall risks that attack underlying systems of verification, knowledge production, and communication, in addition to risks that undermine individual rights.

For example, with regard to organizational risk, we'll see many private companies fail, due to AI product innovation or AI-infused business models. And, of course, organization failures will have a material impact on the individuals connected to the organization - e.g., shareholders, employees, upstream suppliers, customers, etc. These are the kind of systemic risks that will need to be accounted for in a robust risk management scheme.

Section III. Normative Frameworks: The Current Ones and Our Proposal

A review of the governance frameworks that are beginning to emerge around the globe reveal that the first and most important question to answer is what normative framework should guide the approach to governance. In general, countries are selecting normative frameworks that extend the existing normative frameworks shaping governance within that national jurisdiction. In this section, we review some of the most prominent normative frameworks for AI governance in use around the globe. Then we recommend an alternative normative framework.

Existing Normative Frameworks Operating Around the Globe

China: In July 2023, key central government ministries and agencies including the Cyberspace Administration of China ("CAC") published the "Interim Measures for the Management of Generative Artificial Intelligence Services," which came into effect on August 15, 2023 (Zhang, 2023). The measures outline China's regulatory goals for generative AI, which are to ensure its responsible growth and standardized application, while also "safeguard[ing] national security and social public interests" and "protect[ing] the legitimate rights and interests of citizens, legal persons, and other organizations" (Baughman, 2023). As per Article 4 of the Interim Measures, the provision and use of generative AI services must must reflect socialist core values and not contain or generate content that subverts state power, challenges the socialist system, causes harm to national image, undermines social stability, or upsets economic and social order.

Compared to a previous draft issued in April 2023, the Interim Measures aimed to foster a more supportive environment for commercial and research initiatives. The updated version removed a section pertaining to initially strict and onerous obligations on service providers, which would make companies responsible for AI system outputs, for checking the legitimacy of the source of any training data, and for users to register under real identities (Huang, Toner, Haluza, Creemers, Webster 2023; Zhang 2023). Instead, the Interim Measures in Article 7 requires service providers to "take effective measures

to improve the quality of training data, and enhance the authenticity, accuracy, objectivity, and diversity of training data." (Baughman, 2023).

Previous regulations include rules over conspicuous labels on synthetically generated content and the prohibition of the algorithmic generation of fake news (Creemer and Webster, 2022; Creemer, Webster and Toner, 2022). These regulations are significant in their explicit focus on content generation, monitoring, and control.

European Union: Passed on December 8, 2023, the EU's "AI Act" puts an emphasis on risk-based tiering and penalties for non-compliance (Lynch, 2023). The EU's summary of the Act designates its risk categories as follows:

> [The Act] assigns applications of AI to three risk categories. First, applications and systems that create an unacceptable risk, such as government-run social scoring of the type used in China, are banned. Second, high-risk applications, such as a CV-scanning tool that ranks job applicants, are subject to specific legal requirements. Lastly, applications not explicitly banned or listed as high-risk are largely left unregulated. (EU Commission Online, 2023).

The AI Act complements existing legislation aimed at safeguarding individual rights, such as the General Data Protection Regulation, applicable as of 25 May 2018 across all EU member states. For instance, Article 22 of the GDPR declares that individuals shall have the "right not to be subject to a decision based solely on automated processing, including profiling" (GDPR, 2018).

Japan: Japan's strategies and regulations are strongly influenced by its project announced in 2016, "Society 5.0," which aims to resolve social problems such as the aging population through technological innovations[5] (Cabinet Office's Council for Science, Technology, and Innovation, 2016). In March 2019, the Integrated Innovation Strategy Promotion Council published "Social Principles of Human-Centric AI" as principles for integrating AI in society around three core topics: human dignity, diversity and inclusion, and sustainability. The document further outlines seven social principles of AI-capable society: "(1) human-centric[ity], (2) education/literacy, (3) data protection, (4) ensuring safety, (5) fair competition, (6) fairness, accountability and transparency, and (7) innovation" (Social Principles of Human-centric AI).

In outlining these principles, Japan has focused not on restricting the use of AI to protect these principles but rather to bolster them through AI, centering on the upsides of AI's positive impact on society (Habuka, 2023). Currently, Japan does not have any regulations that directly restricts the use of AI. A whitepaper published by the Ministry of Economy, Trade, and Industry (METI) in July 2021 states that "legally-binding horizontal requirements for AI systems are deemed unnecessary at the moment" (METI, 2021). As of September 2023, the Japanese government has partnered with a number of big technology firms such as NEC, Fujitsu, and SoftBank to create LLMs centered around the intricacies of the Japanese language, for instance, including expressions of politeness and cultural appropriateness (Hornyak, 2023).

United Kingdom: The UK places a premium on precedent and common law evolution of policy, including in AI governance. In a 2023 White Paper from the Department for Science, Innovation and Technology, the UK government reiterates its "strong approach to the rule of law, supported by [its] technology-neutral legislation and regulations" and claims that its existing laws cover many ever-growing risks posed by AI technologies such as discrimination, product safety, and consumer law (Secretary of State for Science, Innovation and Technology, 2023). The White Paper proposes that the UK rely on existing regulators such as the Health and Safety Executive, the Equality and Human Rights

Commission, and the Competition and Markets Authority to regulate AI in their respective industries, as opposed to introducing a new commissioner or regulatory body to regulate AI.

United States: On October 10, 2023, President Joe Biden issued the "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" prioritizing a broad conception of safety and a focus on innovation and competition, supplemented by rights and equity commitments to protect consumers, workers, and small businesses (EO 14110, 2023). This framework reinforces a dual focus on national security and global economic competitiveness, which have been the twin peaks of U.S. policy for seventy years. The executive order primarily relies on the existing structure of U.S. agencies, while also establishing multiple coordinating vehicles anchored in different locations from the White House to the Department of Commerce.

From this cursory review of jurisdictionally specific strategies we can see patterns that reflect previous historical directions. The EU, UK, and Japan introduced liberal, rights-protective frameworks, with varying degrees of emphasis on individual rights and societal goals. The U.S. also embraces a liberal, rights-protective framework, but with added attention to marginalized populations and equity, as well as a high priority on national security and economic competitiveness. China explicitly frames its policy around socialist values, giving no attention to the protection of rights. Yet all three approaches share an interest in political stability and in blocking ethnic hatred and discrimination, at least in the letter of the law.[6] Nonetheless, those shared commitments, and adherence to them, are likely to be the necessary building blocks of global governance regimes.

Major transformations of socio-economic relationships need to be navigated with a consistent values-based vision. The goals of protecting privacy, accountability, and transparency are insufficient as guides for the present moment because they do not in themselves include a governance vision. They give us a framework for thinking about how to protect individual rights but little guidance for how society should steer through organizational, systemic, and/or existential risks and opportunities. They support reactive and punitive approaches to governance but no vision for how to construe the risks and opportunities much more broadly, and to make correspondingly large investments in public goods and personnel. In addition, the governance regimes adopted by the world's democracies will require a still fuller normative framework. They will require frameworks that also consider how to protect political rights and rights of participation beyond the existing regulatory apparatus. A new vision for AI governance should, therefore, not only encompass the protection of individual rights but also proactively shape opportunities for the public good and societal well-being.

Proposed Normative Framework for Governing Technology

Our alternative normative framework incorporates key aspects of the liberal framework but reaches farther toward support for public goods. Importantly, it starts from the question of how technology can best advance human flourishing and draws on a tradition of egalitarian pluralism.

To date, too much of our governance of technology, as well as the various stages of research, development, and deployment, have been left in the hands of profit-maximizing firms. When product design teams gather at the whiteboard in big-tech office parks and startup garages, they ask themselves: How do we make a useful product for people? How do we keep our customers in our ecosystem and maintain/create a competitive advantage? Is our product better than our competitors? How can we make money? But one question is rarely asked: Does our technology advance human flourishing?

This is an essential question. By now, we are accustomed to the ambitious mission statements of tech companies: to "organize the world's information," or "make the world more open and connected," on a scale of billions of human beings. Over the last few years, we've also become desensitized to the

gap between those ideals and some of the darker consequences of technology for democracy, civil society, and even conversations at the dinner table supporting diverse viewpoints (Zalesne and Pyati, 2023). How will our technology serve us better if we aren't asking whether it advances human flourishing? We believe that this question should be at the center of technology governance. As public sector leaders face the challenge of governing new AI technologies, this is the question they should keep front and center, and require technology companies to answer responsibly. Technology companies should be responsible to all the stakeholders they affect, not just shareholders. While the insistence that consumers "want" something may justify technologies that bring only minimal risks, it cannot suffice as a guide to the development and deployment of high and unacceptable risk technologies.

We propose an initial overarching normative proposition (proposition 1), and three corollary normative propositions (propositions 2-4), as a framework for governing technology.

Proposition 1: Technology, properly conceived, ought to advance human flourishing.
This assertion, that the purpose of technology is to advance human flourishing, is both familiar and radical. Tech companies' lofty mission statements suggest that technology is an unqualified force for progress and the improvement of the human condition. As the Japanese governance framework articulates, new AI technologies should be human-centric.

Yet, companies are profit-maximizing, and their commitments to maximize shareholder value do not always translate to maximizing human flourishing. For instance, profit-seeking led to pollution of water systems by chemical companies, before regulators stepped in to require them to internalize in their business models what had been negative externalities. This isn't always a case of market failure per se; at times, the markets are poorly designed. For example, many foundation models today are trained off data scraped from an information commons over which data creators lack residual economic or governance rights. Or the pursuit of profit through attention-maximizing algorithms has driven negative mental health outcomes and the creation of dangerous local news deserts.

To assert unequivocally that human flourishing is the purpose of technology is to acknowledge the innovation role of the private sector, but also its limits, especially when new markets are being created that may lack competitive market mechanisms, and are therefore vulnerable to capture or domination by a few over many (Zalesne and Pyati, 2023; Allen et al., 2023; Allen et al., 2022a).

This point is foundational. Society can't steer the technology that shapes our lives if we don't first declare its purpose. In this century, the ambition of tech companies is matched by their proven ability to alter every facet of life. With this power, they *will* accelerate our flourishing or our degradation. It is an indispensable first step to say out loud that we prefer to flourish.

Proposition 2: Human flourishing requires individual autonomy.
Human flourishing flows from individual autonomy. Such autonomy includes both negative liberties, where we are protected in our person, our property, our conscience, our expression, and our associations, and positive liberties, where we govern ourselves in our private lives and share in the governance of our public lives. Ultimately, human beings are creatures who need to chart their own courses in life. We thrive on autonomy, the opportunity for self-creation and self-governance (Pettit, 2014; Allen, 2023). We cannot flourish without it.

The existing paradigms for protecting consumers and human subjects in research and, by extension, for protecting people from harms from AI technologies, tend to be organized around protection of the negative liberties: rights to bodily autonomy, privacy, and non-discrimination with regard to civil, political, and social rights. We have yet to integrate protection for the positive liberties of participation into the basic analysis of risks and opportunities occasioned by novel technologies, and strategies for responding to them.

This insistence on autonomy has obvious geopolitical significance, as autocrats around the world try to strike a bargain with their citizens: prosperity at the price of freedom. The parallels in technology are the products that offer us effortless convenience and pleasing distraction, at the price of our ability to understand what's happening to us or to make a meaningful choice about whether to participate.

The recognition that existing frameworks primarily focus on the negative liberties leads us to articulate the next proposition, which is especially critical for democratic societies, but problematic for non-democratic ones.

Proposition 3: Autonomy requires the values of democratic governance.
The human rights framework is often used to prioritize individual physical and mental safety and integrity and negative liberties. We argue that the positive liberties are equally important to protect—the rights integral to having a role steering one's community and society. These are rights to vote and run for office, and the other elements necessary to give people a chance to see and shape their own communities. Since we live under rules and norms shaped socially, achieving autonomy for individuals requires that they have the chance to participate in shaping the rules and norms that constrain their life. Democracy is necessary for full activation of autonomy. In our recent paper, "Ethics for Decentralized Social Technology," focus on five core values that support democratic governance: (1) difference without domination, (2) individual and community self-determination, (3) egalitarian pluralism, (4) connective and coordinating capacity, and (5) collective ownership of the assets needed for shared governance (Allen et al., 2023). While we won't review these values in detail here, together, they build on lessons from democratic practice to go beyond the surface features of democracy (elections, checks and balances, etc.) to get at the conditions that allow for autonomy and therefore are essential for human flourishing. Governing technology for human flourishing requires keeping these values in mind, too, and steering in the direction of their realization.

Proposition 4: Autonomy requires the material bases of empowerment.
Finally, debates over political economy should consider not just questions of material distribution but also the issue of how economic patterns and institutions affect people's access to empowerment in their lives, communities, and societies (Allen et al., 2022a). A dynamic, inclusive economy, building on the power of the market is critical, but that dynamic market economy needs to integrate all members of society in the productive structure of the economy. This requires steering social transformation toward social connectedness and trust. It requires steering economic transformation toward economic integration and increased power-sharing between workers and holders of capital. Achieving an autonomy-supporting economy also requires steering toward human physical and mental well-being. We need technology that supports these public goods and expands human capacities rather than supplanting the place of human beings in the productive structure of the economy, as we argue in "How AI Fails Us" (Allen et al., 2022a). This is because it is this integration in the productive structure that delivers both material prosperity and empowerment, rather than requiring that people give up the latter for the former.

The assertion here that democratic societies should seek to govern technology so as to steer toward social connectedness, economic integration, and physical and mental health and well-being for residents is not a suggestion that the public sector should take over markets. It's rather to suggest that public investment is needed in public goods that new technologies can support and that the legal frameworks for the operations of market-based firms should drive the pursuit of public goods of these types (Carlin and Bowles, 2021).

Having technology developers themselves embrace this normative framework and seek to develop technologies in these directions would simplify the governance challenge currently presented by new

technologies. That said, the work does not fall to technologists alone. Existing agencies can largely extend pre-existing frameworks for protecting rights and blocking discrimination to the new use cases occasioned by AI. Existing agencies should also consider where and how they are responsible for public goods that advance the normative goals above where new technologies could facilitate provision. And in addition, we will need new capacity to track and steer with regard to the emergence of new capabilities from this point forward.

We have reached the point where we can name and discuss the six governance tasks needed for governance of AI technology.

Section IV. Governance Tasks

The tasks involved in governing AI technologies are: (1) blocking and mitigating harm from both the production and use of AI tools, (2) equipping ourselves to see and maintain human mastery over possibly emergent capabilities, (3) blocking bad actors, (4) steering toward public goods, including through investment in R&D, (5) building human capital; and (6) investing in the sustainability of democratic steering capacity. We will take up each of these governance tasks in turn, reviewing the state of the field and proposing some evolutions in practice.

Before we turn to that task review, however, it is worth noting an important feature of this framework: it integrates attention to near-term already present harms and to existential risk.

Two conceptual frames have come to divide and dominate literature on AI governance. On one hand, philosophers, computer scientists, and industry executives sympathetic to long-term predictions and utility mathematics have taken cues from science fiction to spin out possibilities of extreme and even existential risk from AI systems (Bostrom 2014; CAIS, 2023; Vynck, 2023; Heikkilä, 2023). We critiqued some of these views in "How AI Fails Us" (Allen et al., 2022a). These extreme risks that are predicted share some common features: they will present themselves rapidly, probably snowball, and require significant foresight and planning to avoid them.[7] Although these long-term assessments of possible harm and necessary mitigation efforts do not always invoke the threat of extinction, they tend to prioritize far-out and low likelihood massively scaled catastrophes.

On the other hand, social scientists and students of the history and ethics of artificial intelligence often focus on near-term risks and already present harms, identifying problems such as bias and unfairness in algorithmic design or deployment, or potential violations of privacy. Examples of such present and near-term harms abound in AI ethics and safety literature: biased predictive software perpetuating historical injustices, prejudicial predictive policing, and image-based tools replicating racial prejudices and stereotypes are among the most glaring (Angwin et al., 2016; Simons, 2023; Noble, 2016; Benjamin, 2012; FAccT, 2023). These focuses on present harms have also partially spurred other recent work on datafication, surveillance, and related issues (Valdivia and Tazzioli, 2023; Lazar and Stone, 2023). Other present harms include the unsafe working conditions, climate impacts, and intellectual property violations entailed by the development of novel tools.

Work in AI governance occasionally frames these views in opposition to each other, suggesting that any risk analytic framework must make tradeoffs between a focus on near-term and long-term risk. The distinction between long-term approaches and near-term approaches has become so hegemonic that it pervades the culture of AI governance field work. Policymakers are then left to "balance" what are presented as two fundamentally different worldviews: the present harm approach and the future risk approach. This is not and need not be a Manichean battle, but lines have been drawn in the field, and these views are steadily becoming polarized and indeed politicized (Wong, 2023).

We argue that this supposed tradeoff between governing to address present harms and governing to mitigate future risks, is not a fundamental philosophical problem. It is an organizational problem—that is, a challenge of proper task delegation and governance design. We integrate the management of present harms and potentially emergent risks and opportunities by assigning responsibilities for different regulatory tasks to different public sector agencies. Existing agencies, which can handle near term harms, should be supplemented by a new agency or coordinating task force, capable of building a regulatory structure for management of emergent capabilities.

Both the agencies tasked with addressing near-term harms and those tasked with tracking emergent capabilities, however, should govern not merely in a reactive, status-quo defending way, but with the proactive vision we have sketched above. In both cases, they have the task not merely of properly categorizing and managing narrow risk, but of construing the risks and opportunities much more broadly, in order to make correspondingly large investments in public goods, personnel, and democracy itself.

1. Blocking and Mitigating Harms

The EU AI Act has already gone a long way toward laying out a framework for handling potential harms to individual negative liberties that might ensue from AI technologies. Their strategy has been to focus not on capabilities but on use cases; to regulate the use cases; and to leave the basic technologies themselves fundamentally unregulated. Thus, the EU AI Act introduces a list of high-risk use cases that are permitted subject to compliance with AI requirements and ex-ante assessment, or are subject to information/transparency obligations (AI Act Proposal, 2021). That list includes, among other certain biometric identification and facial-image data scraping, certain uses of predictive algorithms in policing, law enforcement, migration, and judicial processes, and certain uses of AI systems for the determination of eligibility for essential government benefits.

Table 1: EU AI Act Covered Uses (AI Act Proposal, 2021)[8]

1. Biometric identification and categorisation of natural persons:
a. AI systems intended to be used for the 'real-time' and 'post' remote biometric identification of natural persons;
2. Management and operation of critical infrastructure:
a. AI systems intended to be used as safety components in the management and operation of road traffic and the supply of water, gas, heating and electricity.
3. Education and vocational training:
a. AI systems intended to be used for the purpose of determining access or assigning natural persons to educational and vocational training institutions;
b. AI systems intended to be used for the purpose of assessing students in educational and vocational training institutions and for assessing participants in tests commonly required for admission to educational institutions.
4. Employment, workers management and access to self-employment:
a. AI systems intended to be used for recruitment or selection of natural persons, notably for advertising vacancies, screening or filtering applications, evaluating candidates in the course of interviews or tests;

b. AI intended to be used for making decisions on promotion and termination of work-related contractual relationships, for task allocation and for monitoring and evaluating performance and behavior of persons in such relationships.
5. Access to and enjoyment of essential private services and public services and benefits:
a. AI systems intended to be used by public authorities or on behalf of public authorities to evaluate the eligibility of natural persons for public assistance benefits and services, as well as to grant, reduce, revoke, or reclaim such benefits and services;
b. AI systems intended to be used to evaluate the creditworthiness of natural persons or establish their credit score, with the exception of AI systems put into service by small scale providers for their own use;
c. AI systems intended to be used to dispatch, or to establish priority in the dispatching of emergency first response services, including by firefighters and medical aid.
6. Law enforcement:
a. AI systems intended to be used by law enforcement authorities for making individual risk assessments of natural persons in order to assess the risk of a natural person for offending or reoffending or the risk for potential victims of criminal offences;
b. AI systems intended to be used by law enforcement authorities as polygraphs and similar tools or to detect the emotional state of a natural person;
c. AI systems intended to be used by law enforcement authorities to detect deep fakes as referred to in article 52(3);
d. (AI systems intended to be used by law enforcement authorities for evaluation of the reliability of evidence in the course of investigation or prosecution of criminal offences;
e. (AI systems intended to be used by law enforcement authorities for predicting the occurrence or reoccurrence of an actual or potential criminal offence based on profiling of natural persons as referred to in Article 3(4) of Directive (EU) 2016/680 or assessing personality traits and characteristics or past criminal behaviour of natural persons or groups;
f. AI systems intended to be used by law enforcement authorities for profiling of natural persons as referred to in Article 3(4) of Directive (EU) 2016/680 in the course of detection, investigation or prosecution of criminal offences;
g. AI systems intended to be used for crime analytics regarding natural persons, allowing law enforcement authorities to search complex related and unrelated large data sets available in different data sources or in different data formats in order to identify unknown patterns, clusters, or discover hidden relationships in the data.
7. Migration, asylum and border control management:
a. AI systems intended to be used by competent public authorities as polygraphs and similar tools or to detect the emotional state of a natural person;
b. AI systems intended to be used by competent public authorities to assess a risk, including a security risk, a risk of irregular immigration, or a health risk, posed by a natural person who intends to enter or has entered into the territory of a Member State;

c. AI systems intended to be used by competent public authorities for the verification of the authenticity of travel documents and supporting documentation of natural persons and detect non-authentic documents by checking their security features;
d. AI systems intended to assist competent public authorities for the examination of applications for asylum, visa and residence permits and associated complaints with regard to the eligibility of the natural persons applying for a status.
8. Administration of justice and democratic processes:
a. AI systems intended to assist a judicial authority in researching and interpreting facts and the law and in applying the law to a concrete set of facts.

While this is an impressive and comprehensive list, this EU risk framework is still insufficient in three respects. First, it often falls short of accounting for the harms presented by the very development of these technologies even before their use—harms including worker automation, Dickensian labor conditions, and climate impacts. Second, the act fails to identify certain highly dangerous use cases—like facial recognition systems linked to human rights abuses—as worthy of bans and stricter rules (Amnesty, 2023). Third, it falls short of developing a strategy for tracking and pre-empting potentially novel harms that emerge in the event that technological systems become more powerful.

The first two shortcomings can be addressed by using the framework above, while extending it to cover further risks or harms. The third shortcoming, however, requires extending the conceptual architecture of the framework itself. It is in offering this extension that we move beyond the narrow risk-focused treatments of AI governance, to our more expansive and proactive approach to governance, aimed at human flourishing.

The greater challenge comes in recognizing some of the systemic risks that flow from interactions between the technology and other social structures, and so manifest impact on the macro scale. Here regulators need to work more closely with experts in substantive domains of policy to determine whether the right way of mitigating social, economic, health, labor, or educational risks or risks to political stability is through constraints on technology development and deployment or by other means.

We foresee the following interaction risks, listed in *Table 2*, and offer them under three broad categories that align with the normative framework we introduced above, each category tracking one dimension of human flourishing: 1. Individual and community flourishing (consumer protection, user safety, social and mental health, and climate and sustainability); 2. democratic/political stability; and 3. economic empowerment (integration, innovation, and creativity).

Table 2: Risks to Axes and Domains of Flourishing[9]

1. *Individual and Community Flourishing: Consumer Protection, User Safety, Social & Mental Health, Climate & Sustainability*
   a. Societal impact of rapid economic transformation;
   b. Societal impact of potentially shrunken trust and verifiability;
   c. Widening divide between who has access to these types of technologies and tools, and who doesn't;
   d. Environmental risks or factors such as the impacts of mining for rare earth materials (often used in GPUs) and massive energy usage and climate impacts of training and running models (Luccioni et al., 2023);

2. *Harms to Democratic/Political Stability:*
   a. Challenges to political economy and political equality (labor market dislocations and automation, among others);
   b. Epistemic instability due to decreased verifiability of online content
   c. Proliferation of fraud and impacts on the administration of justice and democratic processes.
   d. Expanding power differentials between the public and technology executives and investors;
   e. ncreased opacity of tech-based policy tools, and overreliance on such tools as arbiters of truth and originators of sound decision-making;
   f. Ability of individuals and corporations to take more advantage of jurisdiction surfing; destabilization of domestic legal frameworks in favor of interoperable global regulatory structure;
   g. Increasing likelihood of great power conflict over chips;

3. *Harms to Economic Empowerment (Economic Integration, Innovation, Creativity)*
   a. Labor dislocation in novel sectors, perhaps especially in creative labor markets;
   b. Uncertain and unclear copyright and intellectual property protections;
   c. Labor conditions in mines for metals valued for use in GPUs, in data-labeling and content moderation, and in technology companies;
   d. Exploitative, mercantile practices related to mining and data-scraping harming local economies and violating human rights abroad;
   e. Uncertainty about the future of independent and creative art and cultural products as technologies generate content recombining the work of past artists.

How are the above risks to be addressed? In all cases, models need continuous evaluation for capabilities risk, interaction risk, and alignment during training, pre-deployment, and post-deployment. The work will require participation across the AI value chain, from developers through to firms who may be deploying novel technologies in use cases and applications the developers could never have imagined. Drawing on the EU Risk Categorization Framework, with four levels of risk (unacceptable; high; transparency risk; and minimal or low risk), industry actors and government regulators will need to work together to develop criteria for evaluating which risk category a technology falls within. Standards for evaluation, for external, independent audit of those evaluations, and for required security, should then be developed to align with the risk categorizations. Because these standards for evaluation, audit, and security will need to be put in use across multiple domains and sectors, already subject to existing regulatory bodies, governments will want to form cross-agency learning teams to try to steer toward alignments of conceptualization and vocabulary across context. Such cross-agency learning will in all probability be best advanced by a free-standing AI regulatory body, charged with integrating AI regulation within the procedures of all existing agencies. The U.S. Executive Order does offer examples of precisely such efforts, establishing both a White House AI Council and an interagency AI Council, chaired by the Director of the Office of Management and Budget. Many of the agencies are also charged with creating their own internal, cross-divisional councils. We must also add, though, that while the U.S. Executive Order begins to address risks and opportunities in categories 1 and 3, it overlooks category 2.

For technologies that fall in a minimal or low risk categorization, industry standard setting vehicles and industry validation procedures, supported by private sector third-party auditors, suffice to provide sufficiently protective transparency. While the U.S. Executive Order does charge NIST with standard setting, there is room to supplement that with the development of professional standards by the industry itself. For instance, NIST might work to set up an independent but industry-supported certification board (on the model of the Vitamin Board), from which app developers would seek certification of the safety of their tools to increase the market success of their tools. While labs and app creators would self-declare into the minimal risk category, the catalog of the models and apps certified through the industry board could be routinely audited by the AI regulatory body. This approach might help extend capacity to address a proliferation of tools based on open-source models.

These industry-based methods of certification can be supplemented by standard consumer protection, labor protection, anti-discrimination protections, and health protections, enforced by the relevant federal and state agencies and via litigation, just as the Executive Order proposes. All agencies at both federal and state level will need to develop the capacity to understand how AI tools play a role in the creation of harms to individuals. A good model for this is GAO's Innovation Lab (2023). Every agency will need such an office within it, and every such office should include ethics research capacity. Legislators can and ought to take action to incentivize private developers and technologists to (1) internalize the potential negative externalities of their technologies and (2) develop systems and procedures for public input into and participatory co-design of technological development. Such incentives might take the form of tax or credit program or access to R&D funds (Sitaraman and Narechania, 2023).

Key Recommendations:

1. A focus on individual harm and risk should be widened to include focus on systemic risk, in the categories of public health, social health, and climate and sustainability; democratic/political stability; and economic integration, innovation, and creativity. While some existing frameworks are moving in these directions, all could use further broadening.
2. Existing legal frameworks to protect individual rights and to achieve non-discrimination can function to address many near-term potential harms from AI, but the relevant agencies will require new capacities within their staffs to do this work.
3. A need for cross-domain learning about how best to integrate review and evaluation of AI-based tools and use cases makes it imperative to set up an AI Regulatory body that coordinates across agencies, as the recent U.S. Executive Order and European legislation recognize.

2. Seeing and Mastering Emergent Capabilities

The AI Regulatory Body should also be charged with licensing AI labs conducting research on new models anticipated to introduce new capabilities. In the recent U.S. Executive Order, the Commerce Department is given this role. While it is not granted licensing authority, companies in possession of a certain degree of compute are obliged to register with the Department, via the Defense Production Act. Relatedly, we endorse the decision of the recent U.S. Executive Order to charge the Department of Energy with developing AI model evaluation tools and AI testbeds, as well as with assisting the Department of Commerce "to define, and thereafter update as needed on a regular basis, the set of technical conditions for models and computing clusters that would be subject to the reporting requirements." This is sound assignment given the Department's existing expertise in regulating and monitoring national Labs, as well as the existing integration of the Department of Energy with the National Security infrastructure. The AI Regulatory Body should require that these federally-licensed AI Labs meet well-thought out standards for continuous evaluation of training, pre-deployment decisions, deployment, and security. The AI Regulatory Body should also establish gating criteria for deployment

decisions, when new capabilities are introduced. The AI Regulatory Body will need to be capable of whole systems analysis, and cross-jurisdictional analysis. The Departments of Energy, Defense, and Homeland Security will need to be in a position to conduct independent evaluations of AI technologies. This will require public sector provision of compute sufficient to conduct that work (Schneier and Sanders, 2023). This will require a significant investment of public funds.

Key Recommendations:

1. The proposed cross-agency coordinating AI Regulatory Body (in the U.S. case, the Department of Commerce) should also work with the Departments of Energy and Homeland Security to provide oversight to frontier labs, including via red teaming for emergent capabilities to identify and address emergence of potential catastrophic or existential risks.
2. The Department of Energy should invest in public sector compute, independent from the private sector.

3. Blocking Bad Actors

As many of these novel AI systems are available to the general public, bad actors can use these tools and technologies to advance their nefarious goals. For example, bad actors may use generative AI to develop synthetic content, often called "deep fakes." This altered video, audio or image, is often aimed at leading others to take certain actions or gaining access to sensitive services and data. Deep fakes pollute our information ecosystem, fracturing truth and trust. Some call on citizens to act upon the content (for example, fake declarations by global leaders that may impact global order and lead to armed conflicts). Others impersonate individuals to gain access to sensitive services or data (for example bank accounts); to conduct crimes that rely on "social engineering" techniques (for example phishing and business email compromise attacks). Others might operate in further, perhaps more sophisticated ways.

In addition, bad actors have leveraged advances in AI to design and execute highly sophisticated cyber attacks and hacking, by using AI tools to identify vulnerabilities in other systems, code or even critical infrastructure, and utilizing them to attack those systems. This may create unimaginably severe threats to global security, previously available to nations only. Moreover, AI can assist bad actors in designing highly sophisticated illegal schemes by analyzing systematic failures and global arbitrages. For example, it can be used to design financial crimes such as fraudulent schemes, money laundering, and terrorism financing, by mapping systematic arbitrages in the global financial systems and suggesting ways to launder and funnel illegal funds. Looking forward, these models may aid even more sophisticated schemes as the technology continues to improve.

The White House Executive Order specifically highlights the potential risks of "lowering the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological, or nuclear (CBRN) weapons," cybersecurity attacks, and other unsafe capabilities (EO 14110, 2023). While closed-source models try to get ahead of these use cases by creating boundaries for off-limit topics and capability testing through red-teaming, they have not completely succeeded. The rise in powerful open-source models, and whether public access trumps increased accessibility for bad actors, has been heavily debated. Open-source models make the source code of these AI systems accessible and available to anyone on the internet, therefore bypassing the requirements or accountability of companies who might be held responsible. However, in response to the White House Executive Order, a joint statement released by Mozilla urged that openness will instead improve safety (2023). Despite how bad actors might acquire these systems for their use, it is still critical to develop a response for how we can counteract their nefarious acts.

Both policy-oriented and technically-oriented tools will play a role in blocking bad actors. For example, there are currently tools being developed which identify and mark synthetic content,

authenticate content, and track content origin. Practices such as watermarking, red-teaming, and security testing are becoming more standard across industry as well.

Key Recommendations:

1. Increase transparency and auditability measures in AI systems, in order to understand why and how models are or may be used by bad actors.
2. Advance international cooperation and collaboration on information sharing, security incidents, and ethical frameworks and guidelines.
3. Invest in security protocols, red-teaming, and other security methods to prevent bad actors from manipulating models to extract sensitive information.
4. Re-design the digital ecosystem in ways that will limit the ability to produce synthetic content, this might require an extensive development of tools that can assure authenticity or identify synthetic content and manipulations.

4: Steering toward public goods, including funding R&D

Novel technologies introduce not only risks but opportunities for benefit. Just as we need to analyze novel technologies for their capability and interaction risks, we should also seek to ensure that new opportunities are seized. In some cases, there will be new opportunities for public goods, where commercialization is not the best vehicle for supporting development and scaling of novel technologies. To ensure that we reap not only the private, commercial, and consumer-based benefits of novel technologies, but also public good benefits, we need to see the opportunities new technologies offer to solve collective action problems, and provide public investment for research into and development of those solutions. Just as we considered risks under the guise of three core categories—all aspects of human flourishing and aligned with our normative framework—so too we can look for opportunities via that framework: 1. Individual and community flourishing (consumer protection, user safety, social and mental health, and climate and sustainability); 2. democratic/political stability; and 3. economic empowerment (integration, innovation, and creativity). Now the question to ask is what public goods opportunities are visible along each of these dimensions? *Table 3* addresses some of these opportunities.

Table 3: Opportunities along Axes and Domains of Flourishing

1. *Individual and Community Flourishing:* a. personalization of learning and translation of credentials; education and vocational training b. improved access to expert advice and internet literacy c. contextualization engines to help protect against fraud, misinformation, and disinformation
2. *Democratic/ Political Stability:* a. increased opportunities to engage; b. translation: cross-jurisdictional possibilities;
3. *Innovation, Creation and Economic Integration:* a. improved educational and training opportunities; b. Advances in drug development, cancer research, and other sciences c. entrepreneurial opportunities; d. potentially new jobs emerging; e. “task diversity”—one person can complete many more different kinds of tasks than they could before

We recommend the development of national research and development investment in all of the above public good possibilities. To date, most of the policy frameworks that have been developed are overweighted in their focus on risks. Work remains to be done to identify the many opportunities.

Key Recommendations:

1. AI policy should always attend as much to opportunities as to potential harms.
2. Opportunities should be considered in the domains of public health, social health, and climate and sustainability; democratic/political stability; and economic integration, innovation, and creativity.
3. Pursuing opportunities will require public investment in research and development.

5. Human Capital Strategy

Operationalizing policy recommendations such as these will require significantly increasing the talent pipeline of engineers, scientists, and AI ethicists into public sector service. It will also require greater public education and engagement in ethical questions related to emerging technologies and their potential social impacts. There is an urgent need for colleges and universities with high levels of graduates in technical fields to build an expectation for graduates of national service at some point over the career life course. Among several possible courses of action, the federal and state government could increase investment in public workforce with technical capabilities to monitor the allocation and accumulation of compute power and to evaluate and audit models. Perhaps scholarship and public research funds to this end could connect governments to emerging scholars working on these issues. In addition to technical expertise, government offices may seek greater investment to ensure that government offices and agencies include on their teams people trained to do work on ethics *and* people trained to do work with data and emerging technologies.

Key Recommendations:

1. Colleges and universities with significant proportions of STEM graduates should develop an expectation of national service at some point in the career lifecycle for graduates.
2. AI staff in public sector agencies should always include members trained in ethics. While the opening statement of principles in the U.S. Executive Order includes professionals trained in ethics as among the categories that will be necessary, the full section (Sec. 10) on human capital strategy overlooks this theme (EO 14110, 2023).

6. Investing in the sustainability of democratic steering capacity

Among the most significant challenges presented by the current trend of private, venture-funded technological development is the immense market, platform, and political power placed in the hands of private, profit-motivated individuals. Thus far, our governance framework has aimed to relocate that authority over technological futures from private individuals to public institutions like federal agencies. That task, while important, is only successful if those public institutions are accountable to and steered by a self-determining democratic public. For this reason, we propose a series of reforms aimed at improving the democratic steering capacity of our public institutions. These process-based reforms are both protective and proactive, seeking both to shield important lawmaking systems from capture by technology companies and their tools and to ensure public institutions are and remain accountable to the populations they aim to serve.

Here, we briefly outline a few possible pro-democracy reforms and research agendas that may improve the accountability of and democratic steering mechanisms for regulatory agencies. This is not an exhaustive list, by any means; we merely offer a few examples of the kind of process-based changes

that contribute to more representative institutions capable of legitimately governing technological future and restraining the influence of private actors.

One set of reforms relates to elections and campaigns. In addition to addressing the well-documented issues of dark money in campaigns, voter suppression, and low voter turnout, we ought to focus on one problem that new generative tools may supercharge: the issue of toxic campaign content. Currently, zero-sum opposition campaigns incentivize negative campaigning, large districts increase the distance between Americans and their representatives, and party-based political systems supercharge polarization. Political scientists ought to continue to study potential reforms like state and local implementation of ranked choice voting or all-comers preliminaries instead of party primaries. Reforms such as these (or others) might be avenues to discourage negative campaigning, incentivize coalition formation, and decrease election costs. They might also disincentive and reduce the damage from bots and bad actors flooding information channels with toxic content and misinformation.

Another promising avenue for pro-democratic reform lies in civic education and friendship. So that citizens are equipped to engage in cross-ideological debate over shared technological futures, it may be necessary to invest in our digital literacy and civic mindedness. This might look like expanded civic education programming aimed at developing the kind of "civic friendship" necessary for democratic flourishing (Allen, 2004). As a step in this direction, Congress could pass the Civics Secures Democracy Act, and include funding for civic education providers, and framework developers (for instance, the Educating for American Democracy Roadmap), to integrate education about emerging technology and democracy. Meanwhile, technologists could also build on work on how platforms can better support bridging relationships, instead of division, to incentivize a fresh growth of positive civic culture (Ovadya and Thorburn, 2023).

Finally, the new affordances of AI tools may yield valuable resources for improving the quality of representative governance, as experiments with tools like Pol.is and VTaiwan make clear. The same is true for constituent services, as experiments such as those in Estonia underscore. The challenge of renovating the institutions of democratic representation to serve us well in contexts of significant socio-technical change will require working both on traditional institutional reform and on embedding new digital civic infrastructure within our governing institutions (Allen and Weyl, 2024).

Key Recommendations:

a. We recommend adoption of ranked-choice voting or approval voting systems and non-partisan primaries to provide incentive structures within our systems of representation that can counteract some of the negative externalities of new technologies.
b. We recommend philanthropic and public sector support for investments in embedding new digital civic infrastructure within our governing institutions with the goal of improving the quality of representation.

Section V. Implementation Vehicles

Governance of AI technology in the domestic context can proceed by means of largely familiar vehicles, as is recognized in both the U.S. Executive Order and the EU AI Act. Existing agencies, supplemented by coordinating councils, and reinforced by more staff with increased technical and ethics training, should be able to develop the necessary frameworks. This includes deploying anti-monopoly tools to ensure effective competitive frameworks in support of consumer and public goods and innovation. (Sitaraman and Narachania, 2023). The challenge is to accelerate the pace at which analytical frameworks that clarify the landscape are developed and disseminated. The harder challenges will be to secure global governance frameworks necessary to provide stability for governance in any given domestic context and

to secure governance frameworks for AI companies themselves that tether them to public goods and social accountability.

With regard to the former, an international group with the ability to set and enforce global standards should be formed. We should work to build a global accord around a commitment to (1) no first use of generative foundation models against digital infrastructure or civilians; and (2) no unregulated release to open source. More specifically, policymakers should develop international agencies or organizations for global standards-setting on weaponized AI. An IAEA for AI that would allow for the sovereignty of states over the use of AI within their borders if member states pledge no first use of AI weaponry that targets the civilian population and its related critical infrastructure in other sovereign countries.

With regard to AI companies themselves, the time has come for governance innovation, even experimentation. The transformation of the Open AI nonprofit board underscores the challenge AI labs will have if they try to serve two masters: a nonprofit mission and private, profit-seeking investors. What kind of non-profit board could ensure that an AI Lab, supported by significant private capital, could maintain consistent and reliable adherence to non-profit mission? The requirement to register with the U.S. Department of Commerce and report research breakthroughs there will help. The board will also need to establish strict parameters of investment and tough intellectual property protocols. There might be resources in the tax code as well—and the requirements on nonprofits for the maintenance of their tax exempt status—that could prove a resource to establishing a governance structure for AI Labs that would keep them on mission. Additionally, there might be governance experiments, with representative members of the public added to boards, or routine use of citizen assemblies to support decision-making. But ensuring that this generates a stable regime of incentives will also require a global governance framework that reinforces treatment of AI research in this way.

With respect to the call to create the international framework/task force to lead policy and ensure its implementation by both countries and private sector, a potential model can be the Financial Action Task Force (FATF), an intergovernmental organization that sets global policy on money laundering. A similar framework may be adjusted to the AI domain.

Section VI. Conclusion

Decisions we make in the next few years about how to govern artificial intelligence will be constitutive for much of our economic, social, and political structure. In that context, we need to grapple with some fundamental governance questions. What normative framework should guide us? How do we allocate responsibilities and authorities? How can we ensure that incentive structures reinforce pursuit of public goods? How can we achieve harmonized effort across jurisdictions so that a stable global framework for governing AI comes into existence? Our overarching point is that answering questions of how we should govern this emerging technology is a chance not merely to categorize and manage narrow risk but also to construe the risks and opportunities much more broadly, and to make correspondingly large investments in public goods, personnel, and democracy itself. We have sought to propose initial answers for all of these questions, yet we know that much work remains. Humanity is involved in an era-defining phase of collective learning. We hope only to have contributed to its necessarily incremental advancement and look forward to the further phases of the debate.

Notes

1. See also Executive Order 14110, Sec. 3, for an additional, non-overlapping glossary of key terms (EO 14110, 2023).
2. A helpful survey of some existing definitions of risk in this field comes in Shevlane et al., "Model Evaluation for Extreme Risks," 2023.
3. For a longer, comparative analysis of the climate change and AI risk frameworks, see Novelli et al., 2023.
4. Remco Zwetsloot and Allan Dafoe call structural risk the dimensions of "not only how a technological system can be misused or behave in unintended ways, but also how technology shapes the broader environment in ways that could be disruptive or harmful." The NIST framework also tracks ecosystem harms (e.g., harm to the global financial system or supply chain).
5. Listed innovations include big data, robots, blockchain, and AI. The document further tracks the evolution of human societies from 1.0 (hunter-gatherers), 2.0 (agricultural economies), 3.0 (industrialization), 4.0 (information), up until 5.0 ("super-smart society").
6. This caveat is important. The legally explicit interest in "blocking ethnic hatred" may not always be fulfilled in practice.
7. Milder approaches occasionally leave open the possibility of "extreme" or "transformative" technological change without aiming to predict exactly how new, existential risks might emerge (Acemoglu and Lensman, 2023).
8. This table comes from an earlier version of the EU AI Act. The final agreement text is not yet published, but will be available in the coming months, at which time we expect to update this table with relevant edits.
9. The content of this table draws on several assessments of risks including Shevlane et al., 2022.

References

A Pro-Innovation Approach to AI Regulation. UK Secretary of State for Science, Innovation and Technology. (2023). GOV.UK. Retrieved December 29, 2023, from https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper/.

Acemoglu D., Lensman T. (2023, July 6). Regulating Transformative Technologies. *National Bureau of Economic Research Working Paper Series*. https://ssrn.com/abstract=4512495.

Allen, D. S. (2004). *Talking to Strangers: Anxieties of Citizenship Since Brown v. Board of Education*. University of Chicago Press.

Allen, D. S. (2023). *Justice by Means of Democracy*. The University of Chicago Press.

Allen, D. S. and Weyl, E. G. (2024). AI and Democracy. Forthcoming in *Journal of Democracy*.

Allen, D. S., Benkler, Y., Downey, L., Henderson, R., & Simons, J. (2022). *A Political Economy of Justice*. University of Chicago Press.

Allen, D. S., Frankel E., Lim W., Siddarth D., Simons J., Weyl. E. G. (2023). Ethics of Decentralized Social Technologies: Lessons from Web3, the Fediverse, and Beyond. *Edmond & Lily Safra Center for Ethics*.

Amnesty International. (2023, December 9). EU: Bloc's decision to not ban public mass surveillance in AI Act sets a devastating global precedent. *Amnesty International*. https://www.amnesty.org/en/latest/news/2023/12/eu-blocs-decision-to-not-ban-public-mass-surveillance-in-ai-act-sets-a-devastating-global-precedent/.

Angwin J., Larson J., Mattu S., and Kirchner L., (2016, May 23). Machine Bias. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Artificial Intelligence Risk Management Framework (AI RMF 1.0). (2023, January). *NIST (Department of Commerce)*. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

Baughman J. (2023). Translation: Interim Measures for the Management of Generative Artificial Intelligence Services. *China Aerospace Studies Institute*. https://www.airuniversity.af.edu/Portals/10/CASI/documents/Translations/2023-08-07%20ITOW%20Interim%20Measures%20for%20the%20Management%20of%20Generative%20Artificial%20Intelligence%20Services.pdf.

Beckert, S. (2014). *Empire of Cotton: a Global History* (First edition.). Alfred A. Knopf.

Benjamin, R. (2019). *Race after Technology: Abolitionist Tools for the New Jim Code*. Polity.

Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., Zhang, D., & Liang, P. (2023). The Foundation Model Transparency Index. *arXiv*. https://doi.org/10.48550/arXiv.2310.12941.

Bostrom N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

Bowles S., Carlin W. (2021). Shrinking capitalism: components of a new political economy paradigm. *Oxford Review of Economic Policy.* https://doi.org/10.1093/oxrep/grab029.

Castillo C., Chouldechova A., De-Arteaga M., Ekstrand M., and Lazar S. (2023). Statement on AI Harms and Policy. *ACM FAccT Conference.* Retrieved December 20, 2023, from https://facctconference.org/2023/harm-policy.

Creemer, R., Webster, G., Toner, H. (2022, January 10) Translation: Internet Information Service Algorithmic Recommendation Management Provisions. *DigiChina.* Retrieved December 20, 2023, from https://digichina.stanford.edu/work/translation-internet-information-service-algorithmic-recommendation-management-provisions-effective-march-1-2022/.

Creemers, R. and Webster, G. (2022, February). Translation: Internet Information Service Deep Synthesis Management Provisions (Draft for comment). *DigiChina.* Retrieved December 20, 2023, from https://digichina.stanford .edu/work/translation-internet-information-service-deep-synthesis-management-provisions-draft-for-comment-jan-2022/.

Daníelsson, J., Macrae, R., & Uthemann, A. (2022). Artificial intelligence and systemic risk. *Journal of Banking & Finance*, 140, 106290. https://doi.org/10.1016/j.jbankfin.2021.106290.

De Vynck G. (2023, May 20). The debate over whether AI will destroy us is dividing Silicon Valley. *The Washington Post.* https://www.washingtonpost.com/technology/2023/05/20/ai-existential-risk-debate/.

EU Artificial Intelligence Act Summary (2023). *European Union Artificial Intelligence Act*, retrieved 19 December 2023. https://artificialintelligenceact.eu/.

European Commission (2021, April 21). *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts.* Retrieved https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206.

Europol (2022), Facing reality? Law enforcement and the challenge of deepfakes, an observatory report from the Europol Innovation Lab, Publications Office of the European Union. https://www.europol.europa.eu/publications-events/publications/facing-reality-law-enforcement-and-challenge-of-deepfakes.

Executive Order No. 14110 of Oct. 30, 2023, 88 FR 75191-75226 (2023).

Friedland, A. (2023, May 12). What are generative AI, large language models, and foundation models? *Center for Security and Emerging Technology.* https://cset.georgetown.edu/article/what-are-generative-ai-large-language-models-and-foundation-models/.

General Data Protection Regulation (GDPR)–official legal text. General Data Protection Regulation (GDPR). Retrieved December 29, 2023, from https://gdpr-info.eu/.

Governance Guidelines for Implementation of AI Principles, Ver 1.1. *Ministry of Economy, Trade, and Industry (METI) Expert Group on How AI Principles Should be Implemented (Japan)*, Retrieved from https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20220128_2.pdf.

Habuka H. 2023. (2023, February 14). Japan's Approach to AI Regulation and Its Impact on the 2023 G7 Presidency. Center for Strategic and International Studies. https://www.csis.org/analysis/japans-approach-ai-regulation-and-its-impact-2023-g7-presidency.

Heikkilä M. (2023, June 6). To Avoid AI Doom, Learn from Nuclear Safety. *MIT Technology Review.* https://www.technologyreview.com/2023/06/06/1074077/to-avoid-ai-doom-learn-from-nuclear-safety/.

Henshall, W. (2023, September 19). How China's New AI Rules Could Affect U.S. Companies. *TIME.* https://time.com/6314790/china-ai-regulation-us/.

Hornyak T. (2023, September 14). Why Japan is building its own version of ChatGPT. *Nature.* https://www.nature.com/articles/d41586-023-02868-z.

Huang, S., Toner, H., Haluza, Z., Creemers, R., Webster, G. (2023, April 12) Translation: Measures for the management of generative artificial intelligence services (Draft for comment). *DigiChina.* Retrieved December 19, 2023, from https://digichina.stanford.edu/work/translation-measures-for-the-management-of-generative-artificial-intelligence-services-draft-for-comment-april-2023/.

Joint Statement on AI Safety and Openness. (2023, October 31). Mozilla. https://open.mozilla.org/letter/.

Jones, E. (2023, July 17). Explainer: What is a foundation model? *Ada Lovelace Institute* (blog post). https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/.

Kelley, D. (2023, July 13). WormGPT—The Generative AI Tool Cybercriminals Are Using to Launch Business Email Compromise Attacks. *SlashNext*. https://slashnext.com/blog/wormgpt-the-generative-ai-tool-cybercriminals-are-using-to-launch-business-email-compromise-attacks/.

Lazar, S., and Stone, J. (2023). On the site of predictive justice. *Noûs*, nous.12477. https://doi.org/10.1111/nous.12477.

Lazar, S., and Stone, J. (2023). On the site of Predictive Justice. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 687. https://doi.org/10.1145/3593013.3594035.

Luccioni A. S., Jernite Y., Strubell E. (2023, November 28). Power Hungry Processing: Watts Driving the Cost of AI Deployment? *arXiv*. https://doi.org/10.48550/arXiv.2311.16863.

Lynch, S. (2023). Analyzing the European Union AI Act: What Works, What Needs Improvement. *Stanford University Human-Centered Artificial Intelligence*. https://hai.stanford.edu/news/analyzing-european-union-ai-act-what-works-what-needs-improvement.

Mattu, J. A., Larson, J., Kirchner, L.S. (2016, March 23). *Machine Bias*. ProPublica. Retrieved December 19, 2023, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Morris, I. (2022). *Geography is Destiny: Britain and the World : a 10,000-year History* (First American edition.). Farrar, Straus and Giroux.

Mozur P. (2019, April 14). One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority. *The New York Times*. https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html.

Narechania, T. N., & Sitaraman, G. (2023). An Antimonopoly Approach to Governing Artificial Intelligence. *Vanderbilt Policy Accelerator for Political Economy and Regulation*. https://doi.org/10.2139/ssrn.4597080.

Noble, S. U. (2018). *Algorithms of oppression: How Search Engines Reinforce Racism*. New York University Press.

Novelli, C., Casolari, F., Rotolo, A., Taddeo, M., & Floridi, L. (2023). Taking AI risks seriously: A new assessment model for the AI Act. *AI & SOCIETY*. https://doi.org/10.1007/s00146-023-01723-z.

Office, U.S.G.A. (n.d.). *Science, Technology Assessment, and Analytics (STAA), US GAO*. Retrieved December 29, 2023, from https://www.gao.gov/about/careers/our-teams/STAA.

Ovadya, A., Thorburn, L. (2023) Bridging Systems: Open Problems for Countering Destructive Divisiveness Across Ranking, Recommenders, and Governance. *Knight First Amendment Institute at Harvard University*. https://knightcolumbia.org/content/bridging-systems.

Persily, N., and Tucker, J. A. (2020). *Social Media and Democracy*. Cambridge University Press. https://doi.org/10.1017/9781108890960.

Pettit, P. (2014). Just Freedom: a Moral Compass for a Complex World (First Edition.). W.W. Norton & Company.

Regulatory framework proposal on artificial intelligence. (2023) *EU Commission Online*. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

Sanders, B. S., Nathan E. (2023, December 27). Build AI by the People, for the People. *Foreign Policy*. https://foreignpolicy.com/2023/06/12/ai-regulation-technology-us-china-eu-governance/.

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., Ho, L., Siddarth, D., Shahar Avin, Hawkins, W., Been, K., Iason Gabriel, Bolina, V., Clark, J., Bengio, Y., ... Dafoe, A. (2023). Model evaluation for extreme risks. *arXiv.org*. https://doi.org/10.48550/arxiv.2305.15324.

Siddarth D., Acemoglu D., Allen D., Crawford K., Evans J., Jordan M., Weyl E.g., (2021, December 1). How AI Fails Us. *Justice, Health, and Democracy Impact Initiative*.

Simons, J. (2023). Algorithms for the people: Democracy in the age of AI. *Princeton University Press*.

Social Principles of Human-Centric AI (Japan). Retrieved from https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/humancentricai.pdf.

Society 5.0. (2016). Cabinet Office Home Page. Retrieved December 19, 2023, from https://www8.cao.go.jp/cstp/english/society5_0/index.html.

Statement on AI Risk. *CAIS*. (2023). Retrieved December 29, 2023, from https://www.safe.ai/statement-on-ai-risk.

Understanding and Managing the AI Lifecycle. *GSA* (n.d.). Retrieved December 19, 2023, https://coe.gsa.gov/coe/ai-guide-for-government/understanding-managing-ai-lifecycle/.

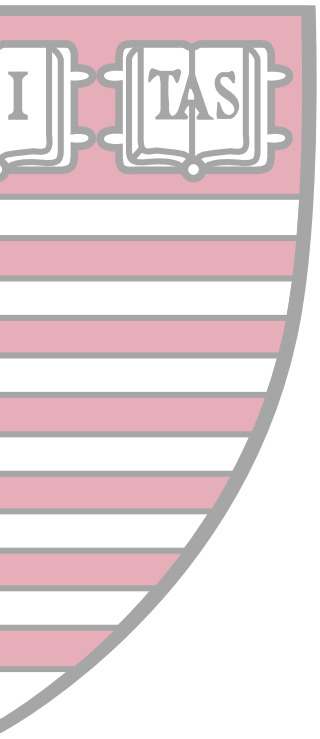
Valdivia, A., and Tazzioli, M. (2023). Datafication genealogies beyond algorithmic fairness: Making up racialised subjects. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 840–850. https://doi.org/10.1145/3593013.3594047.

Wong, M. (2023, June 2). AI Doomerism is a Decoy. *The Atlantic*. https://www.theatlantic.com/technology/archive/2023/06/ai-regulation-sam-altman-bill-gates/674278/.

Zalesne E. K., and Pyati, N. (2023). Putting Flourishing First: Applying Democratic Values to Technology. *Edmond and Lily Safra Center for Ethics.*

Zhang, L. (2023). China: Generative AI Measures Finalized. *Library of Congress* https://www.loc.gov/item/global-legal-monitor/2023-07-18/china-generative-ai-measures-finalized/.

Zwetsloot, R., and Dafoe, A. (2019, February 11). Thinking about Risks from AI: Accidents, Misuse and Structure. *Lawfare*. Retrieved December 19, 2023, from https://www.lawfaremedia.org/article/thinking-about-risks-ai-accidents-misuse-and-structure.



A PUBLICATION OF THE

Ash Center for Democratic Governance and Innovation
Harvard Kennedy School
79 John F. Kennedy Street
Cambridge, MA 02138

617-495-0557
ash.harvard.edu

HARVARD Kennedy School
ASH CENTER
for Democratic Governance
and Innovation