Building a National AI Research Resource:

A Blueprint for the National Research Cloud

WHITE PAPER

Daniel E. Ho
Jennifer King
Russell C. Wald
Christopher Wan

OCTOBER 2021

Stanford University
Human-Centered
Artificial Intelligence

Stanford
Law School

Principal Authors

Daniel E. Ho, J.D., Ph.D., is the William Benjamin Scott and Luna M. Scott Professor of Law, Professor of Political Science, and Senior Fellow at the Stanford Institute for Economic Policy Research at Stanford University. He directs the Regulation, Evaluation, and Governance Lab (RegLab) at Stanford, and is a Faculty Fellow at the Center for Advanced Study in the Behavioral Sciences and Associate Director of the Stanford Institute for Human-Centered Artificial Intelligence (HAI). He received his J.D. from Yale Law School and Ph.D. from Harvard University and clerked for Judge Stephen F. Williams on the U.S. Court of Appeals for the District of Columbia Circuit.

Jennifer King, Ph.D., is the Privacy and Data Policy Fellow at the Stanford HAI. She completed her doctorate in information management and systems (information science) at the University of California, Berkeley School of Information. Prior to joining HAI, she was the Director of Consumer Privacy at the Center for Internet and Society at Stanford Law School from 2018 to 2020.

Russell C. Wald is the Director of Policy for the Stanford HAI, leading the team that advances HAI's engagement with governments and civil society organizations. Since 2013, he has held various government affair roles representing Stanford University. He is a Term Member with the Council on Foreign Relations, Visiting Fellow with the National Security Institute at George Mason University, and a Partner with the Truman National Security Project. He is a graduate of UCLA.

Christopher Wan is a JD/MBA candidate at Stanford University and was teaching assistant for the Stanford Policy Practicum: Creating a National Research Cloud. He also serves as a research assistant for the Stanford HAI and as an investor at Bessemer Venture Partners. He received his B.S. in computer science from Yale University and worked as a software engineer at Facebook and as a venture investor at In-Q-Tel and Tusk Ventures.

The Stanford Institute for Human-Centered Artificial Intelligence
Cordura Hall, 210 Panama Street, Stanford, CA 94305-4101

October 2021, V1.0

Contributors

Many dedicated individuals contributed to this White Paper. To acknowledge these contributions, we list here the contributors for each chapter and section.

Executive Summary and Introduction
Daniel E. Ho, Tina Huang, Jennifer King, Marisa Lowe, Diego Núñez, Russell Wald, Christopher Wan, Daniel Zhang

The Theory for a National Research Cloud
Nathan Calvin, Shushman Choudhury, Tina Huang, Daniel E. Ho, Kanishka Narayan, Diego Núñez, Frieda Rong, Russell Wald, Christopher Wan

Eligibility, Allocation, and Infrastructure for Computing
Daniel E. Ho, Krithika Iyer, Tyler Robbins, Jasmine Shao, Russell Wald, Daniel Zhang

Securing Data Access
Nathan Calvin, Shushman Choudhury, Daniel E. Ho, Ananya Karthik, Jennifer King, Christopher Wan

Organizational Design
Sabina Beleuz, Drew Edwards, Daniel E. Ho, Jennifer King, Christopher Wan

Data Privacy Compliance
Simran Arora, Neel Guha, Jennifer King, Sahaana Suri, Christopher Wan, Sadiki Wiltshire

Technical Privacy and Virtual Data Safe Rooms
Neel Guha, Jennifer King, Christopher Wan

Safeguards for Ethical Research
Daniel E. Ho, Jennifer King, Diego Núñez, Russell Wald, Daniel Zhang

Managing Cybersecurity Risks
Neel Guha, Diego Núñez, Frieda Rong, Russell Wald

Intellectual Property
Sabina Beleuz, Daniel E. Ho, Ananya Karthik, Diego Núñez, Christopher Wan

Case Studies
Daniel E. Ho, Krithika Iyer, Jennifer King, Marisa Lowe, Kanishka Narayan, Tyler Robbins

We also would like to thank Jeanina Casusi, Celia Clark, Shana Lynch, Kaci Peel, Stacy Peña, Mike Sellitto, Eun Sze, and Michi Turner for their help in preparing this White Paper.

External Participants

Guest Lecturers and Interviewees

Erik Brynjolfsson
Stanford University

Isabella Chu
Population Health Sciences
Stanford University

Jack Clark
Anthropic

John Etchemendy
Stanford University

Fei-Fei Li
Stanford University

Marc Groman
Groman Consulting Group LLC

Eric Horvitz
Microsoft

Sara Jordan
The Future of Privacy Forum

Vince Kellen
UC San Diego, CloudBank

Ed Lazowska
University of Washington

Naomi Lefkovitz
National Institute of Standards and Technology (NIST)

Brenda Leong
The Future of Privacy Forum

Amy O'Hara
Georgetown Federal Statistical Research Data Center
Georgetown University

Wade Shen
Actuate Innovation

Suzanne Talon
Compute Canada

Lee Tiedrich
Covington & Burling LLP

Evan White
California Policy Lab
UC Berkeley

In our process, we also engaged many civil society leaders and advocates who have expressed many perspectives about building a National Research Cloud. We have incorporated their feedback where possible and are grateful for their shared thoughts and for helping us shape a better White Paper.

Reviewers

We relied on extraordinary outside expert reviewers for feedback and guidance. We are grateful to Leisel Bogan, Jack Clark, John Etchemendy, Mark Krass, Marietje Schaake, and Christine Tsang for their thoughtful review of the full White Paper, and thank Isabella Chu, Kathleen Creel, Luciana Herman, Sara Jordan, Vince Kellan, Brenda Leong, Ruth Marinshaw, Amy O'Hara, and Lisa Ouellette for their subject expertise on specific chapters.

Workshop Participants

On August 5, 2021, the co-authors hosted a feedback session to hear from a variety of stakeholders in academia, civil society, government, and industry. We are thankful for the time and helpful advice participants offered. Workshop attendee affiliations are listed for identification purposes only. Individuals from Microsoft and AI Now also attended the workshop but did not want to be personally identified.

Taka Ariga
Government Accountability Office

Kathy Baxter
Salesforce

Leisel Bogan
Belfer Center
Harvard University

Jeffrey Brown
IBM

Miles Brundage
Open AI

L. Jean Camp
University of Indiana at Bloomington

Dakota Cary
Center for Security and Emerging Technology
Georgetown University

Shikai Chern
Veritas Technologies

Isabella Chu
Population Health Sciences
Stanford University

Jack Clark
Anthropic

Meaghan English
Patrick J. McGovern Foundation

Cyrus Hodes
The Future Society

Sara Jordan
The Future of Privacy Forum

Vince Kellen
UC San Diego, CloudBank

Michael Kratsios
Scale AI

Samantha Lai
Brookings Institution

Brenda Leong
The Future of Privacy Forum

Ruth Marinshaw
Stanford Research Computing Center

Joshua Meltzer
Brookings Institution

Sam Mulopulous
U.S. Senate

Dewey Murdick
Center for Security and Emerging Technology
Georgetown University

Hodan Omaar
Center for Data Innovation

Calton Pu
Georgia Tech

Asad Ramzanali
U.S. House of Representatives

David Robinson
Upturn

Saiph Savage
Northeastern University

Michael Sellitto
Stanford HAI

Ishan Sharma
Federation of American Scientists

Brittany Smith
Data and Society

John Smith
IBM

Brittany Smith
Data and Society

Victor Storchan
JP Morgan Chase

Keith Strier
AI Compute Task Force Organisation for Economic Co-operation and Development (OECD)

Lee Tiedrich
Covington & Burling LLP

Evan White
California Policy Lab
UC Berkeley

About HAI

Stanford University's Institute for Human-Centered Artificial Intelligence (HAI) applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices. The views expressed in this White Paper reflect the views of the authors.

About the SLS Policy Lab

The Policy Lab at Stanford Law School offers students an immersive experience in finding solutions to some of the world's most pressing issues under the direction of Stanford faculty and researchers. Directed by former SLS Dean Paul Brest, the Policy Lab reflects the school's belief that systematic examination of societal problems, informed by rigorous research, can generate solutions to society's most challenging public problems.

Academic Independence

This White Paper was developed independently by the research team. While we solicited feedback from a wide range of stakeholders, no HAI donors, corporations, or other stakeholders had any involvement with the research and production of this White Paper. Per HAI policy, "Donors cannot dictate research topics pursued by HAI researchers" nor "control permission to publish research results." For more information, please see HAI's policy: https://hai.stanford.edu/about/fundraising-policy.

Disclosures

Stanford University actively engaged and lobbied Congress to pass the *National Artificial Intelligence Research Resource Task Force Act*, working with a coalition of academic, civil society, and industry stakeholders. Co-author Russell Wald provided support to the advocacy efforts.

HAI Co-Director Fei-Fei Li, who served as a guest lecturer in the class, was an early supporter of a task force to study the National Research Cloud. Dr. Li has been appointed to serve as a member of the National Artificial Intelligence Research Resource (NAIRR) Task Force.

Co-author Daniel Ho directs the Stanford RegLab, which has received compute support from HAI's cloud credit program (AWS and GCP), Microsoft's AI for Earth Azure compute credit grant program, and Google's Cloud credit grant for COVID-19 research.

Co-author Jennifer King received unrestricted gift funding for research from Mozilla, Facebook, and Accenture in her previous role at the Center for Internet and Society.

The Stanford Institute for Human-Centered Artificial Intelligence (HAI) receives financial and cloud computing support from A121 Labs, Amazon Web Services, Google, IBM, Microsoft, and OpenAI.

Table of Contents

Case Studies

Executive Summary: Creating a National Research Cloud

Artificial intelligence (AI) appears poised to transform the economy across sectors ranging from healthcare and finance to retail and education. What some have coined the "Fourth Industrial Revolution"[1] is driven by three key trends: greater availability of data, increases in computing power, and improvements to algorithm design. First, increasingly large amounts of data have fueled the ability for computers to learn, such as by training an algorithmic language model on all of Wikipedia.[2] Second, better computational capacity (often termed "compute") and compute capability have enabled researchers to build models that were unimaginable merely 10 years ago, sometimes spanning billions of parameters (an exponential increase in scope from previous models).[3] Third, basic innovations in algorithms are helping scientists to drive forward AI, such as the reinforcement learning techniques that enabled a computer to defeat the world champion in the board game Go.[4]

Historically, partnerships between government(s), universities, and industries have anchored the U.S. innovation ecosystem. The federal government played a critical role in subsidizing basic research, enabling universities to undertake high-risk research that can take decades to commercialize. This approach catalyzed radar technology, the internet, and GPS devices. As the economists Ben Jones and Larry Summers put it, "[e]ven under very conservative assumptions, it is difficult to find an average return below $4 per $1 spent" on innovation, and the social returns might be closer to $20 for every dollar spent.[5] Industry in turn, scales and commercializes applications.

CHALLENGES TO THE AI INNOVATION ECOSYSTEM

Yet this innovation ecosystem faces serious potential challenges. Computing power has become critical for the advancement of AI, but the high cost of compute has placed cutting-edge AI research in a position accessible only to key industry players and a handful of elite universities.[6] Access to data—the raw ingredients used to train most AI models—is increasingly limited to the private sector and large platforms[7], since government data sources remain largely inaccessible to the AI research community.[8] As the National Security Commission on AI (NSCAI) has determined, "[t]he consolidation of the AI industry threatens U.S. technological competitiveness."[9]

Four interrelated challenges illustrate this finding: First, we are seeing a significant brain drain of researchers departing universities.[10] In 2011, AI Ph.D.s were roughly as likely to go into industry as academia.[11] Ten years later, two-thirds of AI Ph.D.s go into industry, and less than one quarter go into academia.[12] Second, these trends indicate that many university researchers struggle to engage in cutting-edge science, draining the field of the diverse set of research voices that it needs. Third, the fundamental research that would guarantee the United States stays at the helm of AI innovation is being crowded out. By one estimate, 82 percent of algorithms used today originated from federally funded nonprofits and universities, but "U.S. leadership has faded in recent decades."[13] Fourth, government agencies have faced challenges in building compute infrastructure,[14] and there are societal benefits to reducing the cost of core governance functions and improving government's internal capacity to develop, test, and hold AI systems accountable.[15] In short, a growing imbalance in AI innovation tilts toward industry, leaving academic and noncommercial research behind. Given the long-standing role of academic and non-commercial research in innovation, this shift has substantial negative consequences for the American research ecosystem.

THE NATIONAL AI RESEARCH RESOURCE TASK FORCE ACT

Responding to these challenges, Congress enacted the National AI Research Resource Task Force Act as part of the National Defense Authorization Act (NDAA) in January 2021.[16] The Act forms part of the National Artificial Intelligence Initiative, which identifies further steps to increase research investments, set technical standards, and build a stronger AI workforce. The Act created a Task Force—the composition of which was announced on June 10, 2021[17]—to study and plan for the implementation of a "National Artificial Intelligence Research Resource" (NAIRR), namely "a system that provides researchers and students across scientific fields and disciplines with access to compute resources, co-located with publicly available, artificial intelligence-ready government and non-government data sets."[18] This research resource has also been referred to as the National Research Cloud (NRC) and was strongly endorsed by the NSCAI, which wrote that the NRC "will strengthen the foundation of American AI innovation by supporting more equitable growth of the field, expanding AI expertise across the country, and applying AI to a broader range of fields."[19]

> The NRC directs more resources toward AI development in the public interest and helps ensure long-term leadership by the United States in the field by supporting the kind of pure, basic research that the private sector cannot undertake alone.

While other initiatives have sought to improve access to compute or data in isolation,[20] the NRC will generate distinct positive externalities by integrating compute and data, the two bottlenecks for high-quality AI research. Specifically, the NRC will provide affordable access to high-end computational resources, large-scale government datasets in a secure cloud environment, and the necessary expertise to benefit from this resource through a close partnership between academia, government, and industry. By expanding access to these critical resources in AI research, the NRC will support basic scientific AI research, the democratization of AI innovation, and the promotion of U.S. leadership in AI.

THEMES

Stanford Law School's Policy Lab program convened a multidisciplinary research team of graduate students, staff, and faculty drawn from Stanford's business, law, and engineering schools to study the feasibility of and considerations for designing the NRC. Over the past six months, this group studied existing models for compute resources and government data, interviewed a wide range of government, computer science, and policy experts, and examined the technical, business, legal, and policy requirements. This White Paper was commissioned by Stanford's Institute for Human-Centered Artificial Intelligence (HAI), which originated the proposal for the NRC in partnership with 21 other research universities.[21]

Throughout our research, we observed three primary themes that cut across all areas of our investigation. We have integrated these themes into each section of our White Paper and drawn on them to explain our findings.

- *Complementarity between compute and data.* As we evaluated the existing computing and data-sharing ecosystems, one of the systemic challenges we observed was a decoupling of compute resources from data infrastructures.

High-performance computing can be useless without data, and a major impediment to data sharing—particularly for high-value government data—lies in requirements for a secure, privacy-protecting computing environment.

> One of the systemic challenges [to basic AI research is] a decoupling of compute resources from data infrastructures. . . . [A] secure, privacy-protecting computing environment [will be critical].

- *Rebalancing AI research toward long-term, academic, and noncommercial research.* Presently, AI innovation is disproportionately dependent on the private sector. Public investment in basic AI infrastructure can both support innovation in the public interest and complement private innovation efforts. The NRC directs more resources toward AI development in the public interest and helps ensure long-term leadership by the United States in the field by supporting the kind of pure, basic research that the private sector cannot undertake alone.
- *Coordinating short-term and long-term approaches to creating the NRC.* Our research considers many near-term pathways for standing up a working version of the NRC by spelling out how to work within existing constraints. We also identify the structural, legal, and policy challenges to be addressed in the long term for executing the full vision of the NRC.

We summarize our main recommendations here.

COMPUTE MODEL

- **The "Make or Buy" Decision.** The main policy choice will be whether to build public computing infrastructure or purchase services from existing commercial cloud providers.
  - It is well-established that, based solely on hardware costs, it is more cost-effective to own infrastructure when computing demand is close to continuous.[22] The government also has experience building high-performance computing clusters, typically built by contractors and operated by national laboratories.[23] The National Science Foundation (NSF) has also supported many supercomputing initiatives at academic institutions.[24]
  - The main countervailing concerns are that existing commercial cloud providers have software stacks and usability that AI researchers have widely adopted and may consider to be a more user-friendly platform. Commercial cloud providers offer a way to expand capacity expeditiously, although scale and availability will still be constrained by the availability of current graphics processing unit (GPU) computing resources.
  - We recommend a dual investment strategy:
    - First, the compute model of the NRC can be quickly launched by subsidizing and negotiating cloud computing for AI researchers with existing vendors, expanding on existing initiatives like the NSF's CloudBank project.[25]
    - Second, the NRC should invest in a pilot for public infrastructure to assess the ability to provide similar resources in the long run. Such publicly owned infrastructure would still be built under contract or

grant, but could be operated much like national laboratories (e.g., Sandia National Laboratories, Oak Ridge National Laboratory) that own sophisticated supercomputing facilities or academic supercomputing facilities.

- **Researcher Eligibility.** While some have argued the NRC should be open for commercial access, for the purposes of this White Paper, we adhered to the spirit of the legislation forming the NAIRR Task Force and only reviewed the use of an NRC for academic and nonprofit AI research. We recommend that the NRC eligibility start with academics who hold "Principal Investigator" (PI) status (i.e., most faculty) at U.S. colleges and universities, as well as "Affiliated Government Agencies" willing to contribute previously unreleased, high-value datasets to the NRC in return for subsidized compute resources. PI status should be interpreted expansively to encompass all fields of AI application. Students working with PIs should presumptively gain access to the NRC. Scaling the NRC to meet the demand of all students in the United States may be challenging, but we also recommend the creation of educational programs as part of the new resource to help train the next generation of AI researchers.
- **Mechanism.** In order to keep the award processing costs down, we recommend a base level of compute access to meet the majority of researcher computing needs. Base-level access avoids high overhead for grant administration and may meet the compute demands for the supermajority of researchers. For researchers with exceptional needs, we recommend a streamlined grant process for additional compute access.

DATA ACCESS MODEL

- **Focus on Government Data.** We focus our recommendations for data provision/access to government data because: (1) there are already a wide range of platforms for sharing private data,[26] and (2) distribution by the NRC of private datasets would raise a tangle of thorny IP issues. We recommend that researchers be allowed to compute on any datasets they themselves contribute, provided they certify they have the rights to that data, and the use of such data is for academic research purposes.
- **Tiered Access.** We recommend a tiered access model: By default, researchers will gain access to government data that is already public; researchers can then apply through a streamlined process to gain access at higher security levels on a project-specific basis. It will be critical for the NRC to ultimately displace the current fragmented, agency-by-agency relational approach. By providing secure virtual environments and harmonizing security standards (e.g., Federal Risk and Authorization Management Program (FedRAMP)[27]), the NRC can collaborate with proposals for a National Secure Data Service[28] to provide a model for accelerating AI research, while protecting data privacy and prioritizing data security.
- **Agency Incentives.** To incentivize federal agencies to share data with the NRC and improve the state of public sector technology, we recommend the NRC permit federal agency staff to use the NRC's compute resources. In keeping with the practices of existing data-sharing programs, such as the Coleridge Initiative,[29] we also recommend that the NRC provide training and support to work with agencies to modernize and harmonize their data standards.
- **Strategic Investment for Data Sources.** In the short term, we recommend that the NRC focus its efforts on making available non-sensitive, low- to moderate-risk government datasets, rather than sensitive government data (e.g., data about individuals) or data from the private sector, due to data privacy and intellectual property concerns. Researchers can still use NRC compute resources on private data but should rely on existing mechanisms to acquire data for their own private buckets on the NRC. For example, images taken from Earth observation satellites, such as

Landsat imagery, provide a promising low-risk, high-reward government dataset, as making such satellite imagery freely available to researchers has generated an estimated $3-4 billion in annual economic benefits, particularly when combined with high-performance computing.[30] Agencies such as the National Oceanic and Atmospheric Administration, the U.S. Geological Survey, the Census Bureau, the Administrative Office of the U.S. Courts, and the Bureau of Labor Statistics, for instance, also have rich datasets that can more readily be deployed. In the long run, access to high-risk datasets, such as those owned by the Internal Revenue Service (IRS) and the Department of Veterans Affairs (VA), will depend on the tiered access model.

ORGANIZATIONAL FORM

Where to institutionally locate the NRC poses a tradeoff between ease of coordination to obtain compute and ease of data access. For instance, locating the NRC within a single agency would make coordination with compute providers easier, but would make data access across agencies more difficult, absent further statutory authority. Many efforts to make data access to government data easier, most notably the Foundations for Evidence-Based Policymaking Act of 2018, have proven to be among the most daunting challenges of government modernization.[31] Building on those insights, we ultimately recommend that the NRC be instituted as a Federally Funded Research and Development Center (FFRDC) in the short run, and a public-private partnership (PPP) in the long run.

- **FFRDC.** FFRDCs at Affiliated Government Agencies would reduce the significant costs of securing data from those host agencies. This approach will also cohere with the greater reliance on commercial cloud credits in the short run, making compute and data coordination less central. In the long run, however, streamlined coordination between data and compute may be more difficult with FFRDCs hosted at specific agencies when (1) the NRC moves away from commercial cloud credits and toward its own high-performance computing cluster, and (2) a greater number of interagency datasets become available.
- **PPP.** In the long run, we recommend the creation of a PPP model, governed by officers from Affiliated Government Agencies, academic researchers, and representatives from the technology sector, which can house both compute and data resources.

ADDITIONAL CONSIDERATIONS

- **Data Privacy.** As an initial matter, an NRC where sensitive or individually identifiable administrative data from multiple agencies are used to build and train AI models will face challenges from the Privacy Act of 1974.[32] The Act is intended to put a check on interagency data-sharing and disclosure of sensitive data without consent.
  - In order to avoid conflicts with nonconsensual interagency data-sharing, we recommend that the NRC should not be instituted as its own federal agency, nor should federal agency staff be allowed access to interagency data.
  - To avoid conflicts with the Act's "no disclosure without consent" requirement, any data released to the NRC must not be individually identifiable. Despite these constraints, the majority of AI research will likely fall under the Act's statistical research exception, contingent on proposals aligning with an agency's core purpose.
  - Given concerns about the potential privacy risks, federal agencies may desire to share data, contingent on the use of technical privacy measures (e.g., differential privacy). While useful in many instances, technical

approaches are no panacea and should not substitute for data access policies.

- The NRC should explore the design of virtual "data safe rooms" that enable researchers to access data in a secure, monitored, and cloud-based environment.
- Additional legislative interventions could also facilitate data-sharing with the NRC (e.g., requiring IT modernization to include data-sharing plans with the NRC).

- **Ethics.** Rapid innovation in AI research raises a host of potential ethical challenges. Given the scope of the NRC, it will not be feasible to review every single research proposal for potential ethical violations, particularly since ethical standards are still in flux. The NRC should adopt a twofold approach.
  - First, for default PI access to base-level data and compute, the NRC should establish an ex-post review process for allegations of ethical research violations. Access may be revoked when research is shown to manifestly and seriously violate ethical standards. We emphasize that the high standard for a violation should be informed by the academic speech implications and potential political consequences of government involvement in administering the NRC and determining academic research directions.
  - Second, for applications requesting access to restricted datasets or resources beyond default compute, which will necessarily undergo some review, researchers should be required to provide an ethics impact statement. One of the advantages of beginning with PIs is that university faculty are accountable under existing IRBs for human subjects research, as well as to the tenets of peer review.
  - We urge non-NRC parties (e.g., universities) to explore a range of measures to address ethical concerns in AI compute (e.g., an ethics review process[33] or embedding ethicists in projects[34]).
- **Security.** We recommend that the NRC take the lead in setting security classifications and protocols, in part to counteract a balkanized security system across federal agencies that would stymie the ability to host datasets. The NRC should use dedicated security staff to work with Affiliated Government Agencies and university representatives to harmonize and modernize agency security standards.
- **Intellectual Property (IP).** While the evidence on optimal IP incentives for innovation is mixed, we recommend that the NRC adopt the same approach to allocating patent rights, copyrights, and data rights to NRC users that apply to federal funding agreements. The NRC should additionally consider conditions for requiring NRC researchers to disclose or share their research outputs under an open-access license.
- **Human Resources.** Given its ambition, significant human resources—from systems engineers to data officers, and from grants administrators to privacy, ethics, and cybersecurity staff—will be necessary to make the NRC a success.

Given its ambition, significant human resources—from systems engineers to data officers, and from grants administrators to privacy, ethics, and cybersecurity staff—will be necessary to make the NRC a success.

Introduction

In March 2020, Stanford's Institute for Human-Centered Artificial Intelligence (HAI) published an open letter, co-signed by the presidents and provosts of 22 top universities in the country, to the president of the United States and U.S. Congress urging adoption of a National Research Cloud (NRC).[1] The NRC proposal aims to close a significant gap in access to computing and data that, proponents argue, has distorted the long-term trajectory of artificial intelligence (AI) research.[2] Without access to such critical resources, AI research may be dominated by short-term commercial interests and undermine the historical innovation ecosystem where basic, fundamental, and noncommercial research have laid the foundations for applications that may be decades away, not yet marketable, or promote the public interest.

In January 2021, Congress enacted the National Artificial Intelligence Research Resource Task Force Act (NAIRR), constituting a task force to consider the design of the NRC.[3] The task force was announced in June of this year and includes one of the original proponents of the NRC and co-director of HAI (Fei-Fei Li).[4]

This White Paper is the culmination of a two-quarter, independent policy practicum at Stanford Law School's Policy Lab program, which was co-taught by three of us (Ho, King, Wald) and a teaching assistant (Wan) and brought together law, business, and engineering students to contemplate key design dimensions of the NRC. We interviewed and convened a wide range of stakeholders, including privacy attorneys, cloud computing technologists, government data experts, cybersecurity professionals, potential users, and public interest groups. Students researched governing legal provisions, policy options, and avenues for the institutional design of the NRC. The practicum team worked independently to shape its recommendations.

The proposal for an NRC is an ambitious one, and this White Paper covers a lot of ground. We begin with the fundamental question—*why* build the NRC (Chapter 1)?—and spell out what we view as a cogent theory of impact. We then cover *who* should have access to the NRC (Chapter 2), *what* comprises the NRC (Chapter 2), *how* access to restricted data may (or may not) be granted (Chapter 3), and *where* the NRC should be located (Chapter 4). We spend extensive time on the data access portion (Chapters 3, 5, and 6), due to the complexities of government data-sharing under the Privacy Act of 1974.[5] As we note in those chapters, the data portion of the NRC is complementary to long-standing efforts to enable greater research access to administrative data under, for instance, the Foundations for Evidence-Based Policymaking Act of 2018[6] and the National Secure Data Service Act proposal.[7] Such sharing must be carried out securely and in a privacy-protecting fashion. We also consider questions of ethical standards (Chapter 7), cybersecurity (Chapter 8), and intellectual property (Chapter 9) that inform the design of the NRC.

We recognize the complexity of the enterprise and that there are many questions not answered herein. The contemplated scale of the NRC may be to AI what the Human Genome Project was to genomics (or what particle accelerators were to physics): public investment for ambitious, noncommercial fundamental scientific research to ensure the long-term flourishing of a critical area of innovation for the United States. There are many areas where we wish we had had the opportunity to engage in more extensive research. We hope this White Paper nonetheless, will provide a useful contribution for the NAIRR Task Force, Congress, the White House, and all those interested in the AI innovation ecosystem.

We owe gratitude to the many people who contributed time, feedback, and insights. Most importantly, we thank the extraordinary students who shaped this White Paper: Simran Arora, Sabina Beleuz, Nathan Calvin, Shushman Choudhury, Drew Edwards, Neel Guha, Krithika Iyer, Ananya Karthik, Kanishka Narayan, Tyler Robbins, Frieda Rong, Jasmine Shao, and Sadiki Wiltshire. We benefited from too many individuals to name, but special thanks go to Taka Ariga, Kathy Baxter, Miles Brundage, Jean Camp, Shikai Chern, Bella Chu, Jack Clark, Kathleen Creel, John Etchemendy, Deep Ganguli, Eric Horvitz, Sara Jordan, Vince Kellen, Mark Krass, Sebastien Krier, Ed Lazowska, Brenda Leong, Fei-Fei Li, Ruth Marinshaw, Michelle Mello, Amy O'Hara, Hodan Omaar, Saiph Savage, Marietje Schaake, Mike Sellitto, Wade Shen, Keith Strier, Suzanne Talon, Lee Tiedrich, Christine Tsang, and Evan White for helpful insights and feedback. HAI staff and research assistants who were essential in helping us during the final stages of editing and compiling the White Paper include Tina Huang, Marisa Lowe, Diego Núñez, and Daniel Zhang.

As we spell out in this White Paper, the NRC is an idea worth taking seriously. It is worth being clear, however, what it would and would not solve. The NRC *would* enable much greater access to—and in that sense, democratize—forms of AI and AI research that have increased in computational demands, but it would *not* categorically prevent or shift the centralization of power within the tech industry. The NRC *would* shift the attention of current AI efforts into more public and socially driven dimensions by providing access to previously restricted government datasets, addressing longstanding efforts to improve access to high-value public sector data, but it would *not* create a system to prevent all unethical uses of AI. The NRC *would* facilitate audits of large-scale models, datasets, and AI systems for privacy violations and bias, but it would *not* be tantamount to a regulatory requirement for fairness assessments and accountability. It is neither a tool of antitrust nor a certification body for ethical algorithms, which are areas worth taking seriously in independent policy proposals.[8] These broader considerations, however, do play into key areas of design and have very much informed our recommendations below on the design of the NRC.

While it alone cannot solve all that ails AI, the NRC promises to take a major affirmative step forward.

Chapter 1: The Theory for a National Research Cloud

This chapter articulates a theory of impact for the NRC. In conventional policy analytic terms,[1] what problem (or market failure) does the NRC address? From one perspective, AI innovation is vibrant in the United States, with major advances occurring in language, vision, and structured data and applications developing across all sectors. Yet from another perspective, current commercialization of past innovation masks systematic underinvestment in basic, noncommercial AI research that could ensure the long-term health of technological innovation in this country.

Current commercialization of past innovation masks systematic underinvestment in basic, noncommercial AI research that could ensure the long-term health of technological innovation in this country.

Our case for the NRC is grounded in both efficiency and distributive rationales. First, the NRC may yield positive externalities, particularly over time, by supporting investments in basic research that may be commercialized decades later. Second, it may help to level the playing field by broadening researcher access to both compute and data, ensuring that AI research is feasible for not just the most elite academic institutions or large technology firms. Given the scale of economic transformation AI is posited to initiate over the next few decades, the stakes are potentially significant. While the largest private interests like platform technology companies and certain elite academic institutions continue to design, develop, and deploy AI systems that can be readily commercialized, a different story is playing out for the public sector and the vast majority of academic institutions, which lack access to core inputs of AI research. The rising costs associated with carrying out research and development are exacerbating the disconnect between current winners and losers in the AI space.

This chapter proceeds in three parts. First, we survey the current landscape of AI research. Second, we articulate shifting trends in AI research and the academic-industry balance. Third, we spell out the risks of federal inaction and the benefits to an investment strategy that couples data and compute resources.

KEY TAKEAWAYS

- The federal government will play a central role in shaping, coordinating, and enabling the development of AI.
- AI research and development is increasingly dependent on access to large-scale compute and data, causing migration of AI talent from the academic to private sector and limiting the range of voices able to contribute to AI research.
- Noncommercial and basic AI research is critical to the long-term health of the innovation ecosystem.
- An NRC that provides data and compute access will help to promote the long-term national health of the AI ecosystem and mitigate the risks of widening inequalities in the nation's AI landscape.

THE AI RESEARCH LANDSCAPE

The field of AI research, as we consider it in this White Paper, is broadly construed. It includes not only academics who identify themselves as researchers in artificial intelligence or machine learning, but also the broader community of researchers who use applied AI in their work, as well as those who examine its impacts on society and the environment.

Many believe, consistent with the legislation calling for the NAIRR Task Force, that AI will have a dramatic impact on society. Nine of the world's 10 current largest companies by market capitalization are technology companies that place AI at the core of their business models.[2] Recent figures from the AI Index demonstrate the growing amount of investment AI companies have drawn. The most recent 2021 iteration of the Index details how global private investment in AI has grown by 40 percent since 2019 to a total of $67.9 billion, with the United States alone accounting for over $23.6 billion.[3] While multiple private sector predictions of the economic impact of AI emphasize the potential for AI to drive significant economic growth through a strong increase in labor productivity, others worry about the pace of structural change in the labor market and economic dislocation for workers automated out of their jobs or impacted by the gig economy.[4]

Such impacts are expected across domains. AI holds substantial promise to transform healthcare and scientific research: AI-related progress in the field of protein folding is poised to dramatically expedite vaccine development and pharmaceutical drug development.[5] The integration of AI-related systems into agriculture may improve crop yields through targeted use of pesticides and soil monitoring.[6] And national security experts have identified AI as a key driver of novel defense capabilities,[7] including cyberwarfare and intelligence collection.

Many countries have recognized the significance of AI as a driver of progress in economic, scientific, and national security, releasing national plans coordinating investment for continued progress in AI.[8] China's national plan announced billions of dollars in funding aimed at making the country the global leader in AI by 2030.[9] The Japanese government partnered with Fujitsu to build the world's fastest supercomputer (Fugaku).[10] Compute Canada has similarly provided research computing access to academics across the country. The U.K.'s national high-end computing resource, HECToR, was launched in 2007 at a cost of $118 million and used by nearly 2,500 researchers from more than 250 separate organizations who produced over 800 academic publications.[11]

The U.S. government initially presented a more decentralized approach, providing support for AI development through National Science Foundation grants and defense spending, but refrained from releasing a unified national plan to coordinate resources across government, private industry, and universities.[12] The creation of a National AI Initiative Office,[13] the updating of the National Strategic Computing Initiative,[14] and the release of the National Security Commission on Artificial Intelligence's (NSCAI) final report[15] introduced a more comprehensive and coordinated approach. Within the United States, the closest model to the NRC may be the COVID-19 HPC consortium, which quickly provisioned compute of 50K GPUs and 6.8 million cores for close to 100 projects across 43 academic, industry, and federal government consortium members united by the common goal of combating the COVID-19 pandemic.[16]

Historically, partnerships between government, universities, and industry have anchored the U.S. innovation ecosystem. The federal government played critical roles in subsidizing basic research, enabling universities to undertake high-risk research that can take decades to commercialize. This approach catalyzed radar technology,[17] the internet, [18] and GPS devices.[19] This history informed the NSCAI's recommendation for substantial new investments in AI R&D by establishing a national AI research infrastructure that democratizes access to the resources that fuel AI. Many policymakers believe that substantial investment will be needed over the next several years to support these efforts, while returns on such investments could potentially transform America's economy, society, and national security.[20]

To be sure, some may challenge the theory of impact. First, some studies dispute the premise that AI will be economically transformative. Some economists argue that

many of the optimistic assessments fail to consider how constrained the uptake of AI innovation may be due to AI's inability to change essential yet hard-to-improve tasks.[21] Others similarly critique the evidence for a fourth industrial revolution.[22] Second, some suggest that the provisioning of the NRC may strengthen the position of large platform technology companies (which of course provokes debates over antitrust in the technology sector[23]), as the NRC may be hard to launch without some involvement of hardware or cloud providers in the procurement process. Third, some would argue that the NRC would generate large negative externalities in the form of energy footprints. For instance, one study found that the amount of energy needed to train GPT-3, a leading natural language processing (NLP) model, required the greenhouse emissions equivalent of 552.1 tons of carbon dioxide,[24] approximately 35 times the yearly emissions of an average American.[25] Expanding access to compute without appropriate controls may contribute to wasteful computing.[26] Finally, some critics argue that any advances in AI are inherently too risky for further investment,[27] given widely documented risks of bias,[28] unintended consequences,[29] and harm.[30]

We are cognizant of these critiques and take them seriously. This White Paper proceeds on the operative premise animating the NRC legislation: that it will be important for the country to maintain leadership in AI—including rigorous interrogation of its uses, limits, and promises—and that this requires supporting access to compute and data. Public investment in AI research for noncommercial purposes may help to address some of the issues of social harm we see presently in commercial contexts[31], as well as contribute to shifting the broader focus of the field toward technology developed in the public interest by the public sector and civil society, including academia. The preceding considerations, however, have shaped our views in key respects, such as the sequential investment strategy, given the uncertainty of AI's potential; the serious consideration of publicly owned infrastructure; the provisions for ethical review of compute and data access; and, most importantly, the enablement of independent academic inquiry into the potential harms of AI systems. The NRC is not an endorsement of blind and naïve AI adoption across the board; it is a mechanism to ensure that a greater range of voices will have access to the basic elements of AI research.

The NRC is not an endorsement of blind and naïve AI adoption across the board; it is a mechanism to ensure that a greater range of voices will have access to the basic elements of AI research.

SHIFTING SOURCES OF AI RESEARCH

We now articulate how and why AI research has migrated away from basic, long-term research into commercial, short-term applications.

First, many current advances fueled by large-scale models are costly to train, relative to the size of typical academic budgets. For example, the estimated cost of training Alphabet subsidiary DeepMind's AlphaGo Zero algorithm, capable of beating the human world champion of the game Go, was more than $25 million.[32] For reference, the total annual 2019 budget for Carnegie Mellon University's Robotics Institute, one of the premier academic research institutions in the nation, was $90 million.[33] A white paper from the Bipartisan Policy Center[34] and the Center for a New American Security noted that the FY2020 budget for non-defense AI R&D announced by the White House was $973 million. In contrast, the combined spending on R&D in 2018 by five of the major technology platform companies was $80 billion. In sum, research universities cannot keep pace with private sector resources for compute. This is not to say that large-scale compute is necessary for all academic AI research, or that academic research is in competition with industry research, but it does illustrate why certain sectors of AI research are no longer accessible to the academic researcher.

Second, the academic-industry divide masks significant disparities between academic institutions. Using the QS World University Rankings since 2012, Fortune 500 technology companies and the top 50 universities have published five times more papers annually per AI conference than universities ranked between 200 and 500.[35] Private firms also collaborate six times more with top 50 universities than with those ranked between 301 and 500.[36] This internal compute divide across universities poses significant challenges for who is at the table.

Third, basic AI research has lost human capital.[37] When this is combined with decreased access to compute and data in the academy, the prospect of conducting basic research at universities becomes less attractive. Top talent in AI now commands private sector salaries far in excess of academic salaries.[38] The departure of AI faculty from American universities has led to what some analysts have dubbed the AI Brain Drain: While AI Ph.D.s in 2011 were roughly as likely to go into industry as academia, two- thirds of AI Ph.D.s now go into industry and less than a quarter go into academia.[39] One study suggests that the departure of AI faculty also has a negative effect on startup formation by students.[40]

While AI Ph.D.s in 2011 were roughly as likely to go into industry as academia, two-thirds of AI Ph.D.s now go into industry and less than a quarter go into academia.

Fourth, as large-scale AI research migrates to industry, the focus of research inevitably shifts. While academic researchers in AI may lack access to the volume of data needed to train AI models,[41] large-platform companies have access to vast datasets, including those about or created by their customers. This data divide in turn distorts AI research toward applications that are focused on private profit, rather than public benefit.[42] Put more colorfully by Jeff Hammerbacher, "The best minds of my generation are thinking about how to make people click ads."[43] The NRC can play a key role in unlocking access to public sector data, which may help to reorient the focus of AI research away from private sector datasets.[44]

The hollowing out of academic AI capacity can be seen in OpenAI's analysis of the relationship between compute and 15 relatively well-known "breakthroughs" in AI between 2012 and 2018.[45] Although the analysis was meant to emphasize the role of computing power, it also illustrates an emerging gap between private sector and academic contributions over time. Of the 15 developments examined, 11 were achieved by private companies while only four came from academic institutions. Furthermore, this imbalance increases over time: Though private sector research has continued accelerating since 2012, academic output has stagnated. The last of the major compute-intensive breakthroughs in OpenAI's analysis stemming from academia was Oxford's 2014 release of its VGG image-recognition program; NYU's work on Convolutional Neural Networks dates back to 2013. From 2015 to 2018, all eight breakthroughs included in OpenAI's analysis came out of private companies. Taken together, this leads observers to argue that academic researchers are increasingly unable to compete at the frontier of AI research.[46] While academic researchers have continued to make important contributions in AI, these are increasingly restricted to less compute-intensive problems. With fewer compute-intensive academic breakthroughs, AI innovations have focused on private interests (e.g., online advertising) as opposed to long-term, noncommercial benefits. To be sure, the private sector has, of course, been central to AI research, but the concern is about the long-term balance of the AI innovation ecosystem.

SCOPING FEDERAL INTERVENTION IN DATA AND COMPUTE

How can we achieve a more balanced approach toward research and development? We first consider the risks of federal inaction and discuss some of the unique advantages of addressing data and compute together.

Risks of Federal Inaction

The risks of federal inaction are twofold. First, basic AI research that has to date paved the way for advances in AI and machine learning will slow. According to a recent study, approximately 82 percent of the algorithms used today originated from nonprofit groups and universities supported by government spending.[47] Even when industry research is successful, it is typically product-focused or incremental, harder to reproduce, and may not be published or open-sourced. An interesting case lies in recent breakthroughs in protein folding. In late 2020, the Alphabet subsidiary DeepMind announced that it had developed a program called AlphaFold, an AI-driven system capable of accurately predicting the structure of a vast number of proteins, using only the sequence of nucleotides contained in its DNA. Whether out of concern for the privatization or to accelerate adoption of related systems, a consortium of academics, led by scientists at the University of Washington, developed an open source competitor called RoseTTaFold.[48] DeepMind did make AlphaFold available to a broad audience, but the concerns illustrate the risks of science posed by exclusively private AI research, reminiscent of the race to sequence the human genome, where public investment in the Human Genome Project preempted concerns about a private firm patenting the human genome.[49]

Second, federal inaction could widen significant inequalities in the AI landscape. Without increased access to computing, education, and training, large parts of the economy may be unable to adapt—whether in financial services, healthcare, education, or government. Diversifying the range of AI research may also promote progress and productivity. One study suggests that the diversity of AI research trajectories—that is, the specific questions, topics, and problems researchers choose to investigate—has become more constrained in recent years and that private sector AI research is less diverse than academic research.[50] Smaller academic groups with lower private sector collaboration appear to bolster the diversity of AI research.[51] From the standpoint of underdeveloped avenues of research, such as ethics and accountability in AI, increasing the range of research topics and methods in the field raises the likelihood of finding breakthroughs that make additional progress in the long term possible.[52] Recent evidence suggests that between 2005 and 2017, just five metro areas in the U.S. accounted for 90 percent of the growth in innovation sector jobs.[53] According to Stanford economist Erik Brynjolfsson, the likely impact of geographic concentration is "there are a whole lot of people—hundreds of millions in the U.S. and billions worldwide—who could be innovating and who are not because they do not have access to basic computer science skills, or infrastructure, or capital, or even culture and incentives to do so."[54] AI technologies can be hard to diagnose and interpret and be prone to substantial bias.[55] Broadening the set of voices that can interrogate such systems will be critical to an inclusive and equitable future.

In sum, federal investment in public AI infrastructure may promote a more equitable distribution of participation in and gains to AI innovation broadly, bolster U.S. competitiveness, and support fundamental research into noncommercial and public sector applications.

Chapter 2: Eligibility, Allocation, and Infrastructure for Computing

This chapter discusses eligibility, resource allocation, and computing infrastructure for the NRC: *Who* should get access to *what* and *how*?

First, when determining who should get access, it is critical to bear in mind the broad goals of the NRC. As discussed in Chapter 1, there is a large resource gap in academia as compared to private industry. In the interest of supporting basic research and democratizing the field, this section will focus on identifying a target group for eligibility. As we articulate below, we refrain from considering expansion to a broader set of commercial, nonacademic parties because of the NRC's focus on long-term, fundamental scientific research. One of the narrowest approaches would be a specialty faculty model that would target researchers engaged in core AI work. But, the difficulties with defining AI and the rapidly expanding domains in which AI is being applied make this model too constrained to realize the full impact of the NRC. Instead we recommend tracking the most common criterion for federal research funding and advocate that eligibility hinge on "Principal Investigator" (PI) status at U.S. universities.[1] One of the tradeoffs is that PIs may be less diverse than a broader segment of researchers,[2] so a longer-term expansion could consider moving beyond this group. While the NRC aims to train the next generation of AI researchers, we caution that an immediate expansion to all graduate and undergraduate students would pose considerable challenges in scaling. Therefore, we recommend that students primarily gain access by participation in faculty-sponsored AI research, instead of blanket student access, and that they gain training through the creation of educational programs.

Second, we discuss three models for allocating computing credit: development of a new grant process, delegating block compute grants to universities for internal allocation among faculty, or universal access. Each of these models trades off the ease of administration against tailoring for specific NRC goals. We recommend an approach used by other national research clouds—namely a hybrid approach of universal default access for the majority of researchers, with a grant process for excess computing beyond the default allocation. Such an approach would keep administrative costs low for the vast majority of researchers, while enabling tailoring through a competitive grant process for the highest-need users.

KEY TAKEAWAYS

- Researcher eligibility for NRC access should begin with "Principal Investigator" status at U.S. universities.
- The NRC should adopt a hybrid approach of universal default access for the majority of researchers and a grant process when requests for compute or data exceed base levels.
- The NRC should adopt a dual investment strategy by developing programs for expanding access to existing cloud services and piloting the ability to provide publicly owned resources.

Third, we consider the "make-or-buy" decision for the NRC. One option would be for the NRC to provide research grants for the use of commercial cloud services that many researchers already rely on (the "buy" decision). Alternatively, the NRC could create and provision access to a publicly high-performance computing cluster (the "make" decision). It is well-established that, based solely on hardware costs, it is more cost-effective to own infrastructure when computing demand is close to continuous. On the other hand, existing commercial cloud providers have developed highly usable software stacks that AI researchers have widely adopted. Commercial cloud providers offer a way to quickly expand capacity. We hence recommend a dual investment strategy to (a) quickly launch the NRC by subsidizing and negotiating cloud computing for AI researchers with existing vendors, expanding on existing initiatives like the National Science Foundation's CloudBank project; and (b) invest in a pilot for public infrastructure to assess the ability to provide similar resources in the long run. Such publicly owned infrastructure would likely be built under contract or grant, but could be operated much like national laboratories that own sophisticated supercomputing facilities, as is the case with other national research resources (e.g., Compute Canada, Japan's Fugaku).

Our recommendations are informed by a series of case studies that are presented throughout this chapter, as well as through the remainder of the White Paper. Table 1 summarizes how existing models compare on the three key design decisions. At the outset, we note that few existing initiatives have attempted to provide compute power at the scale of the NRC. At the same time, we view the NRC as complementary to more traditional areas of scientific computing.[3]

	ELIGIBILITY			ALLOCATION				OWNERSHIP	
Existing Program	**PI Only**	**Any Faculty**	**Students**	**Existing Grant Process**	**University Allocation**	**New Process**	**Default Access w/Tiers**	**Private**	**Public**
CloudBank	X		X	X					X
Stanford HAI-AWS Cloud Program		X			X			X	
Stanford Sherlock Cluster	X						X	X	
Google Colab		X	X				X	X	
Compute Canada	X						X		X
Fugaku		X				X			X
XSEDE	X	X					X		X
DOE INCITE	X					X			X

Table 1: Key design differences between computing case studies. "Other faculty" indicates an eligibility set for faculty other than PI status (e.g., requiring Stanford affiliation for the Sherlock cluster) and "new process" is used to indicate the creation of any process other than those currently listed (e.g., Fugaku is currently soliciting proposals with research facilities).

Eligibility

The first task is identifying which researchers should be eligible for the NRC. Chapter 1 discussed the need to support AI innovation in universities. Therefore, this section will scope eligibility within academia by analyzing the access-resource trade-offs in alignment with the NRC goals.

At the outset, we note that we do not analyze eligibility in depth beyond academic researchers. The legislation constituting the NRC task force specifically contemplates "access to computing resources for researchers across the country."[4] The NRC is defined as "a system that provides researchers and students across scientific fields and disciplines with access to compute resources."[5] The most natural interpretation of this language suggests a core focus on scientific and academic research.[6]

Introducing commercial access to the NRC, particularly for under-resourced firms such as small businesses and startups, may very well benefit the U.S. innovation ecosystem. But the challenges of incorporating commercial access to the NRC are enormously complex. First, including software developers at startup companies as "researchers" within the meaning of the NDAA would raise a wide range of boundary questions that the NRC may be poorly equipped to adjudicate. According to the Small Business Administration (SBA), there are over 31 million small businesses in the United States.[7] Over 627,000 businesses open each year.[8] Should all such businesses be eligible to compute on the NRC? How would one avoid gaming (e.g., strategic subsidiaries/spinoffs) eligibility? And, how would this advance the scientific mission of the NRC? Second, while potentially valuable, it is less clear how the inclusion of startups and small businesses meets the theory of impact of the NRC. As currently construed, the concern animating the NRC lies in the importance of long-term, noncommercial fundamental research that can ensure AI leadership for decades to come. Commercialization is not the element of the AI innovation ecosystem that faces the structural challenges articulated in Chapter 1. Finally, scaling the NRC to allow meaningful commercial access would pose serious practical challenges. Because the Task Force must also consider the feasibility of the NRC, we have not considered in depth a conception that would extend the term "researcher" to encompass large portions of the commercial private sector. Expansion to non-academic, nonprofit organizations may be a more reasonable consideration, as the objective of some entities (e.g., not-for-profit investigative journalism, civil society organizations) may be closer to the core of the NRC's mission of empowering long-term beneficial research that cannot currently occur.[9] In the long term, the NRC should consider the trade-offs to such an expansion.

Even if the NRC adopts a broader computing model down the road, we believe that focusing on academic researchers is an important starting point as it illuminates some of the main operational considerations for NRC access.

SPECIALTY FACULTY MODEL

One of the narrowest approaches to NRC eligibility would be to restrict it to faculty engaged in AI research. Under this approach, policymakers would direct computing resources exclusively toward faculty working on identifiable AI projects, which often need large amounts of compute power. A benefit of this approach is that researchers' familiarity with the infrastructure would likely mean that fewer funds would be devoted to cloud service training for novice users.

Yet the set of self-identified core AI faculty are few and concentrated in a small number of universities, which are already more likely to gain access to large-scale computing. Limiting access to core AI faculty would hence undermine the mission of democratizing AI research. In addition, the application of AI is expanding rapidly across domains. Interdisciplinary research deploying AI in new domains will be vital for maintaining American leadership in AI, as well as for animating basic research questions. Restricting eligibility to core AI faculty (however defined) could jeopardize the ability of researchers from all academic disciplines (e.g., in the physical sciences, social sciences, and humanities) to contribute to realizing AI's full potential.

GENERAL FACULTY MODEL

A more natural starting point for NRC eligibility is with Principal Investigators (PIs) at U.S. colleges and universities, the most commonly deployed criterion for federal grants. Requirements for PI status are set by individual universities and include a broad range of researchers certified by their university as qualified to lead large research projects.[10] While PI certification may vary from institution to institution, an important baseline criterion of PI status is that the researcher is subject to their institution's training and certification processes, which in turn clarify a researcher's responsibilities regarding the management and execution of their research proposals. Existing programs for allocating computing power typically set eligibility based on PI status as it ensures the researcher has the infrastructure to carry out a large-scale research project. CloudBank, an NSF program that distributes funds for commercial cloud computing resources, awards grants to PIs, who may distribute funds to other researchers and students on the project.[11] Compute Canada allows all faculty granted PI status by their university to automatically receive a preset amount of computing credits and apply for further credit as needed. The PI may then sponsor others to access the credit.[12]

We recognize that PI status does not include all university-affiliated researchers. In 2013, of the over 200,000 self-identified academic researchers, just under 60,000 were employed in a role other than full-time faculty, a position that may not be eligible for PI status.[13] From 1973 to 2013, the percentage of full-time faculty among engineering doctorate holders decreased by 2 percent, while the percentage of "other" academic jobs (including research associates) increased by 12 percent.[14] But, the reliance on PI status would not prevent PIs from allocating access to non-PI status researchers on a project, and administrative ability weighs strongly in favor of consistency with current grant eligibility criteria.

STUDENTS

Should graduate and undergraduate students be able to access the NRC? One of the principal challenges here lies in scale and administrability. One estimate is that there are nearly 20 million college students in the U.S.[15] Second, PI-oriented eligibility does not preclude university students from accessing resources to undertake AI research under the direction of PIs. The Compute Canada model, for example, restricts eligibility to faculty, but allows faculty to sponsor collaborators, including any student researcher. An access model for the NRC that allows PIs to sponsor students provides further research and training opportunities for students. Third, a number of existing cloud services already provide limited access to computing credits for educational purposes. Google Colaboratory, for instance, provides free, but not reliably guaranteed, access to cloud services.[16] Amazon Web Services provides up to $35 of AWS credits for free to all university faculty and students. Despite existing resources, students may need more resources. The Google subsidiary and online community Kaggle, for example, provides 30 hours of GPU access per week for free and found that 15 percent of users exceeded the limit.[17]

While the exact scope of student computing power needs is unclear, we recommend funding an educational resource once researcher needs and resource limitations have been gauged. Currently, the NSF's CloudBank is piloting a Community & Education Resource to earmark a small set of credits for educational purposes.[18] This resource allows a university professor to request a small number of credits for student coursework or small-scale research.

Regardless of which eligibility model the NRC adopts, there will also be a significant need for support staff, training documentation, and educational materials so researchers can effectively make use of the compute and data resources (see Appendix D). The reason some students and researchers may not take advantage of all available cloud credits could, for instance, stem from the difficulty in using cloud platforms. If the NRC serves academics from a range of disciplines, this question of human capital will be especially relevant to serve different models of research. A robust training program for users of the NRC will ensure ease of use and encourage appropriate utilization of the cloud.

Resource Allocation Models

We now consider three resource allocation models: (1) a new grant process; (2) block grant allocation to universities; and (3) universal—but potentially tiered—access.

NRC GRANT PROCESS

Establishing a new grant process for compute access would have one main advantage. The program could be built specifically for the purpose of AI research, with reviewers who are familiar with AI concepts, practices, and trends. Such a process might therefore enable improved allocation decisions and provide the NRC with greater control over its investments.

That said, establishing a peer-review process for all applications would be resource intensive, requiring the establishment of a grant administration program akin to those at the National Science Foundation (NSF) or the National Institutes of Health (NIH). For instance, to implement peer review required for the merit review process, the NSF annually needs a community large enough to conduct nearly 240,000 reviews per year.[19] Since the contemplated reach is broad, we are mindful of adding a significant service burden for faculty conversant in AI for every application for compute access. Peer review for compute access would require significant overhead and delays in compute allocation.

UNIVERSITY ACCESS

To reduce administrative costs, an alternative scheme would be to allocate credits to universities based on the number of eligible researchers. The NRC could allocate resources to universities as block grants, and in turn, rely on the university to distribute computing access. (For example, the NRC could purchase significant amounts of compute from cloud providers, create virtual credits that are convertible into appropriate cloud resources, and delegate allocation to universities.) This approach would have the advantage of tapping into the universities' local expertise for reviewing and distributing resources. It would, however, lead to a highly decentralized process, providing little oversight to understand the distribution of usage, and give the NRC little control over resource allocation. While we do not recommend this route as the principal allocation scheme, we do believe that some allocation to university-based IT support teams may be warranted to support researchers in using the NRC. XSEDE's "Campus Champions" program, for instance, provides university employees access to the system to support the computational transition.[20]

UNIVERSAL ACCESS

The last potential model would provision universal access to base-level compute to all eligible PIs. The closest model is Compute Canada's national research cloud, which provides base-level compute access to all faculty in Canada. This would significantly reduce administrative overhead, both for an institution running the review process, and academics seeking NRC access. The primary downside is that base-level compute may be insufficient for specialized needs.

We recommend combining a universal baseline model with a grant process for compute needs beyond base-level access. The reduced complexity in administering a universal baseline access compute model makes it an attractive option for the NRC in allocating compute resources, especially with respect to the NRC's goal of opening access to compute resources.[21] XSEDE, for instance, uses a similar model of streamlined "Startup Allocations" (issued for one-year terms, typically within two weeks of application) and "Research Allocations" for more significant compute requests. Compute Canada provides access to 15 percent of PIs to increased compute capacity based on a merit competition. A critical question will, of course, be the level of baseline computing that will determine overall costs, physical space requirements, and the like. To benchmark this, we recommend an in-depth study of the anticipated computing needs, based on existing academic computing centers.[22]

The grant process for additional compute could take multiple forms; for example, while one could allow individual PIs to apply directly to the NRC for excess

compute, the NRC could also allocate "blocks" of resources at the university level and allow universities to oversee their administration. In any case, due to the size of such requests, grant reviews should be conducted on a merit basis and administered by a combination of NRC staff and an external advisory board of university faculty. In 2021, Compute Canada, for instance, completed its review of 650 research submissions in about five months, with only 80 volunteer reviewers from Canadian academic institutions to assess the scientific merit of the proposal.[23] In order to avoid conflicts of interest, we strongly recommend against the participation of any faculty or private sector advisers who have conflicts of interest with any vendors that provide services to the NRC. Ideally, proposal review should be independent, blinded, and based on scientific merit to the extent possible.

CASE STUDY: CLOUDBANK

In 2018, the National Science Foundation's (NSF) Directorate for Computer and Information Science and Engineering (CISE) created the Cloud Access Solicitation to provide funding for AI-related research endeavors. Initially created to meet the needs of the NSF funding recipients to access public clouds, CloudBank is an interesting case study for exploring resource allocation models. Accessible through a portal, CloudBank aids researchers in using cloud resources fully by facilitating the process of "managing costs, translating and upgrading computing environments to the cloud, and learning about cloud-based technologies."[24]

CloudBank is a collaboration project established via an NSF Cooperative Agreement with the San Diego Supercomputer Center (SDSC) and the Information Technology Services Division at UC San Diego, the University of Washington eScience Institute, and UC Berkeley's Division of Data Science and Information.[25] Each of these institutions handles an area, according to its comparative advantage.[26] For example, SDSC is responsible for building the online portal, and UC San Diego is in charge of managing the accounts of the users.[27]

CloudBank also aims to reduce the cost of cloud computing: It uses both the ongoing discounts with cloud providers from the University of California and the discounts that come with bulk cloud purchase from the cloud procurement consulting firm Strategic Blue, which regularly partners with the likes of AWS, Microsoft, and Google.[28] Furthermore, there is no overhead cost associated with the cloud allocations through CloudBank, since the terms of the NSF cooperative agreement prohibit indirect costs.[29] With these cost-saving mechanisms, researchers can afford more computing capacities from a variety of major cloud vendors.

By requesting the use of CloudBank during their application to the selected NSF projects,[30] researchers can gain access not only to various advanced hardware resources, but also to a variety of services to make the process more supported and monitored.[31] CloudBank also gives research community members access to its education and training information.[32]

KEY TAKEAWAYS

- ***Built into existing grant process:*** Researchers eligible for certain existing NSF grants can simply request access to CloudBank through the same grant application.
- ***Single point of entry for compute access:*** The CloudBank portal provides a single point of entry for researchers to access funds to use on whichever commercial cloud provider they prefer.
- ***Cost reduction:*** No overhead costs are associated with using CloudBank.
- ***Student access:*** Limited funds are set aside for grants to students and classes.

Computing Infrastructure

Cloud computing environments connect local computing devices such as desktop computers to large, typically geographically distributed servers containing physical hardware. This hardware, in turn, is responsible for storing data and performing computation over computer networks—all of which is mediated through a collection of software services. This model centralizes the usual operational management for those using the network and provides adjustable units of computation and data storage to allow for fluctuations in demand. Users interact with the cloud by launching virtual connections to the server—cloud instances—and running containerized processes remotely. These operations are managed by the cloud and available for monitoring through dashboards. Cloud computing may be serviced through on-premises clusters, via external vendors, or some combination thereof, and accessed over networks with varying security and connectivity, from internet-accessible to air-gapped regions.

The infrastructure to the NRC could be developed with two general approaches: (1) the NRC could use commercial cloud platforms as its infrastructure backbone; or (2) the federal government could engage a contractor to build a high-performance computing (HPC) public facility specifically for the NRC. This section addresses some advantages and disadvantages of both. (We provide an estimated cost comparison of these two approaches in Appendix A.) The two approaches discussed here are not mutually exclusive, and we ultimately recommend a hybrid investment strategy. In the short run, the NRC should scale up cloud credit programs (similar to NSF's CloudBank program) to provide both streamlined base-level access and merit review for applications going beyond base-level access. In the long run, the NRC should invest in a pilot to develop public computing infrastructure. Even with public infrastructure, it will be critical to meet "burst demand" (to expand resources when compute demand peaks). The success of the initial investments should guide the prospective model as to whether to rely on publicly or privately owned infrastructure in the longer term. We note that in order to scale successfully to either resource will require building institutional capacity at academic institutions.

COMMERCIAL CLOUD

The greatest advantage of using commercial cloud services for the NRC is that significant infrastructure already exists.[33] Under this model, the NRC would simply subsidize credits for using commercial cloud services (similar to NSF's CloudBank program). Thus, rather than spending years building new computing resources, policymakers could launch the NRC soon after they determine the program's administrative details. (We note, however, that there may still be significant GPU shortages in the short run; with the contemplated scale of the NRC, significant infrastructure would need to be built.) Since many researchers already use commercial cloud services for their AI research, the transition into the NRC program could be relatively seamless. Furthermore, commercial cloud platforms offer the NRC greater flexibility to change the size and scope of the program. Commercial cloud platforms charge for the amount of compute actually used.[34] Thus, the size of the NRC could expand or retract in line with shifting demand. In contrast, a dedicated HPC system would have a set amount of hardware that costs the same, no matter how effectively it's being used.

Working directly with commercial cloud providers also offers several advantages for the NRC. The commercial cloud services market is highly competitive and features numerous providers capable of meeting the NRC's needs. The NRC would have the option of using one provider or multiple providers. If opting to use just one provider, the government's bargaining power may be at its strongest in helping to drive down prices for the NRC. Alternatively, using multiple providers gives the NRC greater flexibility in available services and hardware. Either way, policymakers would have the opportunity to negotiate contracts and prices with commercial cloud providers every few years, which will be critical to cost containment.[35] The NRC would also not be locked into using the same provider or set of providers for the duration of the program. Rather, NRC staff could reevaluate which commercial cloud provider's infrastructure would best meet the NRC's needs at the start of each new contract.

CASE STUDY: XSEDE

The Extreme Science and Engineering Discovery Environment (XSEDE) is an NSF-funded organization that integrates and coordinates the sharing of advanced digital services such as supercomputers and high-end visualization and data analysis resources.[36] XSEDE is a collaborative partnership of 19 institutions, or "Service Providers," many of which are nonprofits or supercomputing centers at universities and provide computing facilities for XSEDE researchers.[37] XSEDE supports work from a wide variety of fields, including the physical sciences, life sciences, engineering, social sciences, the humanities, and the arts.[38] XSEDE allocations are available to any researcher or educator at a U.S. academic, nonprofit research, or educational institution, not including students.[39] However, researchers can share their allocations by establishing user accounts with other collaborators, including students.[40]

Researchers have two different paths to requesting allocations: Startup Allocation and Research Allocation. Startup Allocations apportion XSEDE resources for small-scale computational activities.[41] Startup Allocations are one of the fastest ways to gain access to and start using XSEDE resources, as requests are typically reviewed and awarded within two weeks.[42] Startup Allocation requests also require minimal documentation: the project's abstract and the researchers' curriculum vitae (CV).[43] Startup Allocations typically last for one year, but requests supported by merit-reviewed grants can ask for allocations that last up to three years. Researchers can also submit renewal requests if their work needs ongoing low-level resources.[44]

For research needs that go beyond the computational limits under a Startup Allocation, researchers must submit a Research Allocation request.[45] XSEDE strongly encourages its users to request a Startup Allocation prior to requesting a Research Allocation, in order to obtain benchmark results and more accurately document their research needs in the Research Allocation.[46] Research Allocation requests must include a host of documents, such as a resource-use plan, a progress report, code performance calculations, CVs, and references.[47] Requests are accepted and reviewed quarterly by the XSEDE Resource Allocations Committee (XRAC), which assesses the proposals' appropriateness of methodology, appropriateness of research plan, efficient use of resources, and intellectual merit.[48]

XSEDE abides by a "one-project rule," whereby each researcher only has one XSEDE allocation for their research activities.[49] For instance, if a researcher has several grants that require computational support, those lines of work should be combined into a single allocation request. This minimizes the effort required by the researcher to submit requests and reduces the overhead in reviewing those requests.

XSEDE also uses a "Campus Champion Program" to streamline access to resources.[50] The Campus Champion Program is a group of over 700 Campus Champions who are employees or affiliates at over 300 U.S. colleges, universities, and research-focused institutions.[51] These Campus Champions facilitate and support use of XSEDE-allocated resources by researchers, educators, and students on their campuses. For instance, the Campus Champions host awareness sessions and training workshops for their institutions' researchers while also capturing information on problems and challenges that need to be addressed by XSEDE resource owners.[52]

Finally, XSEDE welcomes collaboration opportunities with other members of the research and scientific community.[53] For example, XSEDE assists other organizations in acquiring and operating computing resources and helps to allocate and manage access to those resources. Recently, XSEDE worked with academics and private industry to form the COVID-19 High Performance Computing Consortium, which provides researchers with powerful computing resources to better understand COVID-19 and develop treatments to address infections.[54]

KEY TAKEAWAYS

- ***Federally-funded infrastructure:*** XSEDE is an NSF-funded initiative that integrates and coordinates shared supercomputing and data analysis resources with researchers.
- ***Tiered access to compute:*** For baseline access to compute, XSEDE leverages a fast, low-hurdle review process. For access beyond the baseline, XSEDE has its own resource allocations committee that reviews applications every quarter.
- ***"Campus Champions Program:"*** XSEDE partners with employees and affiliates at colleges, universities, and research institutions to help researchers get access to compute resources.
- ***Collaboration:*** XSEDE collaborates with the private sector in acquiring, operating, and managing compute resources.

Commercial cloud platforms also provide other advantages to the NRC. The labor of managing, maintaining, and upgrading the hardware behind the NRC would be handled by private parties that already have expertise in running cloud services at scale and have invested billions of dollars into doing it. This arrangement allows researchers access to a greater variety of hardware that is constantly being expanded and upgraded.[55] With a strong economic incentive to keep improving cloud offerings, commercial cloud services offer an assortment of instance types—i.e., the various permutations and combinations of GPU/CPU, memory, storage, and networking specifications that constitute a compute instance—with different hardware at a range of price points. Thus, researchers would have the flexibility to choose both what hardware would best fit the needs of their projects and how best to allocate their limited cloud credits. Researchers could also have access to cutting-edge technology specially designed for AI research, such as chips optimized for training and inference, developed and exclusively used by commercial cloud providers.

Using commercial cloud services for the NRC comes with significant tradeoffs, however. While the initial costs of subsidizing cloud credits might be less than building public infrastructure, many studies show that relying on commercial cloud services would likely be much more expensive in the long term.[56] For example, a study of Purdue University's Community Cluster Program shows that the amortized cost of its on-premises cluster over five years is 2.73 times cheaper than using AWS, 3.24 times cheaper than using Azure, and 5.54 times cheaper than using Google Cloud.[57] A similar study at Indiana University estimates that the total investment into its locally owned supercomputer, Big Red II, is about $10.1 million, while the total cost of a three-year reservation on AWS about $24.9 million.[58] Cost comparisons in other studies are even more dramatic. For instance, a study of the Advanced Research Computing clusters at Virginia Tech shows that the five-year cost for its on-premises cloud is about $15.5 million, while the five-year cost for reserved AWS instances using the same workloads would be about $136.3 million.[59]

What explains these cost disparities? Estimates comparing commercial cloud services to a dedicated HPC cluster show that commercial cloud services are more expensive per compute cycle.[60] At least in part, this is due to the fact that commercial services are optimized for commercial applications. Compute Canada, for example, found that building their own infrastructure was cheaper than using commercial services, because they did not have the same core use needs as commercial customers, a tradeoff that gained their system more computing power at the expense of availability.[61] Although the analysis was published in 2016, Compute Canada's own benchmarking of costs concluded:

While the initial costs of subsidizing cloud credits might be less than building public infrastructure, many studies show that relying on commercial cloud services would likely be much more expensive in the long term.

> Currently, it is far more cost effective for the Compute Canada federation to procure and operate in-house cyberinfrastructure than to outsource to commercial cloud providers. . . . Cloud-based costs ranged from 4x to 10x more than the cost of owning and operating our own clusters. Some components were dramatically more expensive, notably persistent storage which was 40x the cost of Compute Canada's storage.[62]

Ultimately, the cost difference between commercial cloud services and HPC systems depends on how often and how efficiently the HPC system is used. We provide a cost calculation that updates Compute Canada's below, arriving at cost differentials of comparable magnitude. Commercial cloud instances with comparable hardware under constant usage, even with substantial discounts,

would be significantly more expensive over time for the NRC than a dedicated HPC system. Bringing the cost of commercial cloud services under that of an HPC system would require policymakers to either negotiate exceptionally high discounts with commercial cloud providers or make major sacrifices in hardware speed or overall scale of the NRC. A similar cost calculation is also what led Stanford University to simultaneously invest in both on-premises hardware and a commercial cloud-based solution for its Population Health Sciences initiative (see box case study in Chapter 3). The most common practice across NSF centers, such as the XSEDE initiative (see box case study below), is also to build infrastructure instead of relying on commercial cloud credits, due to these cost considerations.

Finally, relying on the commercial cloud may raise questions about industry consolidation. There are two main answers to this question. One is that building a dedicated, publicly owned HPC clusters would require purchasing sophisticated hardware from existing industry players, which also exist in concentrated industries. In other words, it is difficult to imagine no involvement of private industry under either option. Another major constraint lies in time: A fully mature, public infrastructure NRC could not be stood up overnight. Moreover, a publicly owned cloud would still likely require a major technology company to build the infrastructure under contract, as is the case for National Labs, or using a grant, as is the case for XSEDE.

PUBLIC INFRASTRUCTURE

Building a new HPC cluster would be a bespoke solution, tailored to fit the NRC's specific compute needs. This approach would be relatively well-explored territory for the federal government.[63] The U.S. Department of Energy (DOE) and the U.S. Department of Defense (DOD) already regularly contract with a handful of companies to build HPC clusters every few years.[64] The DOE itself already uses two of the three fastest HPC clusters in the world and recently funded the development of two new supercomputers that, when completed, will be the world's fastest by a significant margin.[65] The National Science Foundation commonly issues grants for the construction of high-performance computing infrastructure.[66] Given this familiarity, policymakers would have reasonable estimates for how much a new HPC cluster for the NRC would cost and would already have relationships with the companies that would submit bids for the contract.

The hardware cost for such compute scale are, of course, substantial.[67] For example, the IBM supercomputer used at Oak Ridge National Laboratory (ORNL)—known as "Summit"—cost $200 million.[68] At the time of its completion in 2018, Summit was the fastest supercomputer in the world and, as of 2020, is still the second fastest.[69] Frontier, the new Cray supercomputer being built at ORNL in 2021, cost $500 million. When completed, it is anticipated to be the fastest supercomputer in the world at "up to 50 times" faster than Summit.[70] Nonetheless, these large up-front costs could come with the benefit of computing infrastructure specifically designed for AI research and the NRC's needs. Such a system would be more efficient in cost per cycle over the long term than subsidizing commercial cloud services. The NRC could also expand and upgrade multiple clusters over time to meet the changing needs and scope of the program.

In addition, a dedicated cluster for the NRC has the advantage of giving the federal government greater control over computational resources (e.g., reducing uncertainty over the products and platforms, such as the sudden deprecation of required APIs). This level of control over the hardware also allows policymakers greater flexibility with NRC operations. Taking the public infrastructure approach (i.e., "making" not "buying") comes with several significant trade-offs to weigh against the policy goals of the NRC. First, building a new HPC cluster would take about two years, in addition to the time it takes to solicit and evaluate proposals from potential contractors.[71] If the NRC hopes to quickly stimulate and help democratize AI research in the U.S., such a timeline for the program would not be ideal, given how quickly AI discoveries advance. Of course, contracting with cloud vendors or issuing grants for the construction of supercomputers would also require a process. Yet, building a cluster could raise more challenging contracting issues, such as budget overruns and project delays.[72] Contractors' experience with building this type of hardware may help mitigate some of these concerns, as well as their self-interest in being

considered for future government contracts. But the risks are nonetheless still present.

Second, the usability and the feature set of the software stack for public infrastructure is by no means proven. One of the most common hurdles to researcher adoption of cloud computing lies in the usability of systems,[73] and public infrastructure has less of a track record of easing that onboarding path at the contemplated scale. This is why we recommend a pilot to assess whether a national HPC center can be administered in a way to ensure the ease of cloud transition and software stack that researchers have become accustomed with private providers.

Third, policymakers would also need to account for costs of maintaining and administering the system.[74] They would need to find facilities to house and manage the hardware and to account for the high energy costs of running an HPC cluster, as well as disaster prevention and recovery cost.[75] These costs are significant. In 2021, the Oak Ridge Leadership Computing Facility requested $225 million to operate all of its systems.[76] The Argonne Leadership Computing Facility, in turn, requested $155 million.[77] Furthermore, the lifecycle of DOE HPC systems has traditionally been about seven years, after which new systems are built and old ones decommissioned.[78] While it is uncertain what the lifespan of newer systems will be, this seven-year figure would lead us to argue that the NRC should expect to either upgrade its systems or build new ones with some degree of regularity.

Last, giving the federal government greater control over the computing resources would not immediately make the NRC safe from attacks.[79] As with using commercial cloud infrastructure, security will primarily be contingent on the NRC's implemented data access model.[80] We discuss security issues in depth in Chapter 8.

CASE STUDY: FUGAKU

In 2014, Japan's Ministry of Education, Culture, Sports, Science and Technology launched a public-private partnership between the government-funded Riken Institute, the Research Organization for Information Science and Technology (RIST), and Fujitsu to create the supercomputer successor to the K computer that supports a wide range of scientific and societal applications.[81] The result was Fugaku, which was named the world's fastest supercomputer in 2020.[82]

The technical aim of Fugaku was to be 100 times faster than the previous K computer, with a performance of 442 petaFLOPS in the TOP500's FP64 high performance LINPACK benchmark.[83] It currently runs 2.9 times faster than the next fastest system (IBM Summit)[84] and is composed of slightly over 150,000 connected CPUs, with each CPU using ARM-licensed computer chips.[85] Despite having around 1.9 times more parts than its K computer predecessor, Fugaku was finished in three fewer months.[86] The six-year budget for Fugaku was around $1 billion.[87]

RIST solicited proposals for usage through the "Program for Promoting Research on the Supercomputer Fugaku." Under the program, Fugaku has already been used to study the effect of masks and respiratory droplets in order to inform Japanese policy during the COVID-19 pandemic.[88] For FY 2021, 74 public and industrial projects were selected for full-scale access to Fugaku.[89] Currently, RIST is still requesting proposals that fall under specific categories of usage, and any interested researcher may apply.[90]

KEY TAKEAWAYS

- ***Significant compute power:*** Fugaku was the fastest supercomputer in 2020.
- ***New application process for compute power:*** Applications were solicited to test out the supercomputer on a host of tasks and have control over who received compute power.

COST COMPARISON

To conclude this chapter, we provide a rough cost comparison between a leading commercial cloud service and a dedicated government HPC system (IBM Summit) (see Appendix A for details). We refer the reader to substantial work that has been published on the economics of cloud computing for a fuller analysis, much of which emphasizes the variance in computing demand.[91]

Building standalone public infrastructure is projected to be less expensive than implementing the NRC through a vendor contracting arrangement over five years. At a 10 percent discount on standard rates over five years, and under constant usage, AWS's more powerful cloud-computing option (known as P3 instances) could cost 7.5 times as much as Summit's total estimated costs, using comparable hardware. We use a 10 percent discount that was negotiated by a major research university with a commercial cloud provider. In contrast, the government would need to negotiate an 88 percent discount for AWS to be cost-competitive with a dedicated HPC cluster in the long run. Even in a scenario where NRC usage fluctuates dramatically, commercial cloud computing could cost 2.8 times Summit's estimated cost. (While variability in usage factors heavily into these estimates, the use of schedulers can contribute to a leveling out of demand.[92])

These cost estimates have important limitations. First, government may be able to negotiate the cost down. We have used as a benchmark one major university's enterprise agreement with AWS, which provides a 10 percent discount, relative to market rates. But, unless the negotiated discount is orders of larger magnitude, the commercial cloud will remain significantly more expensive. Second, these cost estimates primarily focus on computing.[93] As Compute Canada's analysis showed, the cost difference in storage was even greater. Third, the use of commercial rates is likely *more* favorable to cloud vendors, as government security standards typically increase rates due to regulatory requirements. For instance, a "data sovereignty" requirement for data and hardware to reside within the United States, or private cloud requirements for certain agency datasets, may increase the cost of commercial cloud computing significantly. Fourth, this simple cost comparison is static, and does not reflect changes in hardware costs and pricing structures that are likely to occur over a five-year period under rapidly changing market conditions. But, if the NRC in fact scales, systems would be procured incrementally over time, upgrading available resources and providing options at different price points, similar to current commercial options. Last, as noted above, these cost estimates take into account maintenance as budgeted for the Summit, but may not take into account all such non-hardware costs, which is why we recommend a pilot to explore the ability to open up government computing facilities to NRC users.

In short, we offer this simple comparison to highlight some of the salient cost considerations to the make-or-buy decision, which arrives at a very similar conclusion to the analysis done by Compute Canada.

CASE STUDY: COMPUTE CANADA

Compute Canada formed in 2006 as a partnership between Canada's regional academic HPC organizations to share infrastructure across Canada.[94] The organization's stated mission is to "enable excellence in research and innovation for the benefit of Canada by effectively, efficiently, and sustainably deploying a state-of-the-art advanced research computing network supported by world-class expertise."[95]

Compute Canada's infrastructure includes five HPC systems that are hosted at research universities across Canada.[96] From 2015-2019, Compute Canada used about $125 million (CAD) in funding to build four of these systems.[97] They also investigated using commercial cloud resources instead of building these new systems.[98] However, they ultimately concluded that relying on commercial cloud providers would be significantly more expensive and could not provide the desired latency for large-scale, data-intensive research.[99] In 2018, Compute Canada requested $61 million (CAD) to fund its operations, budgeting $41 million (CAD) for operating its HPC systems and $20 million (CAD) on support, training, and outreach.[100] Demand for Compute Canada's HPC resources far exceeds the infrastructure's current capacity and is expected to keep growing.[101] In 2018, Compute Canada estimated they would need about $90 million (CAD) per year over five years to invest in expanding infrastructure to the point where it could meet projected demands.[102]

About 16,000 researchers from all scientific disciplines use Compute Canada's infrastructure to support their work.[103] Compute Canada distributes its resources in two ways. First, Principal Investigators and sponsored users may request a scheduler-unprioritized resource allocation for their research group.[104] Compute Canada finds that many research groups can meet their compute needs this way.[105] Alternatively, researchers who need more or prioritized resources may submit a project proposal to the annual "Research Allocation Competitions."[106] Submitted proposals go through a scientific peer review and a technical staff review to rate their merits.[107] Scientific review examines the scientific excellence and feasibility of the specific research project, the appropriateness of the resources requested to achieve the project's objectives, and the likelihood that the resources requested will be efficiently used.[108] This review is conducted on a volunteer basis by 80 discipline-specific experts from Canadian academic institutions.[109] Technical review is conducted by Compute Canada staff itself, who verify the accuracy of the computational resources needed for each project, based on the technical requirements outlined in the application, and makes recommendations about which resources should be allocated to meet the project's needs.[110] In 2021, Compute Canada received 651 applications to the Research Allocation Competition and fully reviewed all applications in the span of five months.[111]

KEY TAKEAWAYS

- ***Default access with tiers:*** All PIs are eligible for access to a scheduler-unprioritized compute resource allocation with an application process built in for requesting more. Most researchers find the default allocation sufficient for their needs.
- ***Widely used and increasing demand:*** Compute Canada's infrastructure is widely used across academic disciplines, with demand constantly exceeding resources. Compute Canada intends to invest heavily in infrastructure to meet increasing demands.

Chapter 3: Securing Data Access

After compute resources, the next critical design decision for the NRC is how to both store and provide its users access to datasets: the "data access" goal of the NRC. Indeed, as articulated in the original NRC call to action, government agencies should "redouble their efforts to make more and better quality data available for public research at no cost," as it will "fuel" unique breakthroughs in research.[1] Investigating some of the most socially meaningful problems hinges on large but inaccessible datasets in the public sector. From climate data housed by the National Oceanic and Atmospheric Administration (NOAA), health data from the country's largest integrated healthcare system in the Department of Veterans Affairs (VA), or employment data in the Department of Labor (DOL), such data could fuel both fundamental research using AI and refocus efforts away from consumer-focused projects (e.g., optimizing advertising) to more socially pressing topics (e.g., climate change).

As noted in the congressional charge, facilitating broad data access is a crucial pillar of the NRC. Importantly, as we discuss below, we limit the scope of our recommendations to facilitating access to public sector government data, which as a condition of accessing government administrative data, NRC researchers should only use for academic research purposes. NRC users should also be able to compute on any private dataset available to them. There are available mechanisms for sharing such datasets, but we identify the NRC's major challenge as providing access to previously unavailable government data.

Government data is intentionally decentralized. By design of the Privacy Act of 1974, there is no centralized repository for U.S. government data or a core method for linking data across government agencies.[2] The result is a sprawling, decentralized data infrastructure with widely varying levels of funding, expertise, application of standards, and access and sharing of policies. Thus, the NRC will have to develop a unified data strategy that can work with a wide range of agencies, unevenly adopted security standards, and within existing data privacy legislation.

Previous efforts have sought to improve access to and sharing of federal data, both between agencies and with external researchers, but there are still significant barriers to enabling AI research access of the kind that the NRC demands.[3] By linking data governance policies with access to compute, building on existing successful models, and working with agencies to create interoperable systems that satisfy security and privacy concerns, the NRC can enable increased access to data that will aid AI researchers in answering pressing scientific and social questions and increase AI innovation.[4]

KEY TAKEAWAYS

- The NRC should adopt a tiered model for access to and storage of federal agency datasets. Tiers should correspond to the sensitivity of the data.
- The NRC can help to harmonize the fragmented federal data-sharing landscape.
- The NRC should consider incentivizing agency participation by granting agencies that contribute data the right to use NRC compute resources.
- The NRC should strategically sequence data acquisition by focusing first on low- to moderate-risk datasets that are currently inaccessible.
- Due to legal constraints and many outside options, the NRC should focus its efforts on streamlining access to government datasets. Researchers should still be permitted to use NRC compute resources on private datasets, as long as researchers certify they have rights to use such data.

We will first explain why the NRC should focus its efforts on facilitating federal government data sharing rather than private sector data sharing. We then examine how and why the status quo for federal data sharing fails to realize the massive potential of government data. While the concept of centralizing disparate data sources to unlock research insights is not new,[5] there are unique challenges for doing so within the context of the NRC. We will also discuss the key elements of our proposed model: (1) the use of FedRAMP as a system for categorizing datasets based on their sensitivity, and for modifying access to them through tiered credentials for NRC users; (2) promotion of interagency standardization and harmonization efforts to modernize data-sharing practices; and (3) strategic considerations regarding how to sequence efforts in streamlining access to particular datasets.

The case studies included throughout this White Paper were chosen as exemplars of successful data-sharing initiatives[6] and to illustrate the range of available design decisions. While each case study provides a unique glimpse into different approaches, some common themes emerge. First, many of the data-sharing entities we studied not only have a single point of entry for researchers to request access, but also allow government agencies to retain some control over access requirements to their data. As we discuss below, this conception of the NRC as a data intermediary would provide real benefits in streamlining data access while still maintaining trust among agencies that wish to protect their data. Second, some initiatives use funding and personnel training as carrots to incentivize agencies to engage in data sharing. The NRC can learn from these initiatives in formulating its own set of incentives for agencies.

PRIVATE DATA SHARING

Should the NRC affirmatively facilitate private dataset sharing? While there are definite benefits to providing researchers with access to private data,[7] the NRC will have its largest impact by focusing its efforts first on mechanisms to access and share government data.

As an initial matter, a variety of mechanisms for general data sharing already exist.[8] Private sector stakeholders, moreover, can and have often built their own in-house platforms to allow access to approved datasets while minimizing intellectual property concerns,[9] or provide access to their application programming interfaces (APIs) to make open-source data more easily accessible.[10] By focusing on providing access to public sector data, notably administrative data that is traditionally inaccessible to most researchers,[11] the NRC would play a unique and pertinent role for researchers across disciplines without having to deal with complex private-sector data concerns or the need to incentivize participation by nongovernment actors.

Complex intellectual property concerns would arise from the NRC permitting, facilitating, or even requiring private sector stakeholders and independent researchers to share their private data freely alongside public sector data.

Complex intellectual property concerns would arise from the NRC permitting, facilitating, or even requiring, private sector stakeholders and independent researchers to share their private data freely alongside public sector data. First, this would involve complex questions regarding what licenses should be available or mandated for NRC users in order to encourage data sharing, despite apprehensions of how such sharing may affect future profitability and commercialization. While mandating an open-source (e.g., Creative Commons) license would benefit researchers most by providing the broadest access to data and would benefit NRC administrators by removing some possible IP infringement concerns, private sector stakeholders may feel deterred from uploading as a result. Conversely, if users have a choice to adopt a license that

allows them to preserve their IP rights, private sector stakeholders may feel more comfortable sharing their data, but this would shift some liability to users—or to the NRC itself—by relying on users to abide by the license. This would involve an emphasis on enforcement, ranging from explanations and user disclaimers to the industry standard of a full-blown notice-and-takedown system.

Data owners may want to prevent the uploading of copyrighted works by, for instance, having the NRC itself assess whether private data is already protected by copyright. Industry standards for conducting data diligence, using manual or automated tools, would either be very labor intensive[12] or prohibitively expensive.[13] Even if these industry standards were met, researchers may find an NRC data-sharing platform duplicative.

None of the above would prevent researchers from using NRC *compute* resources on their own private datasets. Like current cloud providers, the NRC can stipulate in an End User Licensing Agreement (EULA) that researchers must agree they own the intellectual property rights on the data they are using.[14] This EULA can also assign liability to the end user, rather than the NRC, for any use of data that is encumbered by existing IP provisions. Additionally, the discussion above pertains to whether researchers should be required to share their *private data*, not to whether researchers should be required to share the *outputs* of their research conducted on the NRC. The latter point is discussed in Chapter 9.

THE CURRENT PATCHWORK SYSTEM FOR ACCESSING FEDERAL DATA

The NRC could play a pivotal role streamlining access to government data in a system that is currently decentralized.[15] In some cases, agencies may simply lack a standardized method for sharing data.[16] Due to perceived legal constraints, risks, or security concerns, agencies often have little practical incentive to share their data.[17] Successful examples of researchers gaining access to government data from individual agencies frequently rely on the researchers having personal relationships with administrators, and a willingness on the part of the administrator to push against these constraints in service of the research project.[18] While this relationship-based process has produced some successes,[19] the far more common outcome is that data is simply not shared or accessed by researchers.[20] Indeed, one government official indicated that overcoming the obstacles to making certain government data available for research was the greatest challenge in a lengthy career.

One government official indicated that overcoming the obstacles to making certain government data available for research was the greatest challenge in a lengthy career.

Agencies typically require the recipient of the data to abide by a data-use agreement (DUA). These DUAs prescribe such limitations on data usage as the duration of use, the purpose of use, and guarantees on the privacy and security of data.[21] However, DUAs suffer from a central problem: The process for negotiating DUAs is highly fragmented and inconsistent across government agencies, drastically increasing the complexity in obtaining approvals for them.[22] Some agencies have a designated office or process to handle DUAs, but other agencies rely on extemporaneous processes and ad hoc, quid pro quo arrangements.[23] One such example is the Research Data Assistance Center, a centralized unit within the Centers for Medicare & Medicaid Services (CMS) dedicated to supporting data access requests.[24] In contrast, DUAs within the Department of Housing and Urban Development and the Department of Education are handled in decentralized business units, each with different routing channels and legal teams, which can confuse reviewers when multiple

data requests between the same parties are routed simultaneously but separately.[25] Indeed, university DUA negotiators in one survey complained that the process was a game of "bureaucratic hot potato" and wondered, "Why isn't there just one template for everything?"[26] Ultimately, the lack of standardization means that DUAs often require extensive review and revision, creating substantial delays.

Agency-by-agency requirements also impede data sharing. These requirements can range from mandating that researchers only access data at an onsite facility, using government-authorized equipment, to capping the amount of computational cycles that can be used to analyze data, or restricting the amount of data available simultaneously.[27] These restrictions are particularly problematic, given that modern AI models can require massive amounts of data and computation to be most effective.

Broadly, the reasons for this dysfunction range from valid concerns about security and liability to the mundane and prosaic. Information technology systems within some agencies operate literally decades behind the technological frontier; a 2016 report from the Government Accountability Office (GAO) detailed examples of these legacy systems, discussing how several agencies were dependent on hardware and software that were no longer updateable, and required specialized staff to maintain.[28] A lack of incentives, a risk-averse culture, and an agency's statutory authority also play an important role in enabling or obstructing data sharing.[29]

We are by no means the first observers to note these problems. Advocates have been working for years to standardize and modernize government practices around data and technology.[30] For example, the Federal Data Strategy is the culmination of a multiyear effort to promulgate uniform data-sharing principles to address the fact that the United States "lacks a robust, integrated approach to using data to deliver on mission, serve the public, and steward resources."[31] However, substantial challenges remain, particularly since the bulk of the efforts focused on opening access to government data have not been undertaken with the specific needs of machine learning and AI in mind.

TIERED DATA ACCESS AND STORAGE

The decentralized nature of government data has cascading implications across many aspects of the government data ecosystem. One key area that will affect the NRC is a lack of consistent storage and authentication access protocols across government agencies.

Because many government datasets contain sensitive data (e.g., high risk due to individual privacy concerns),[32] a crucial component of the NRC's data model will consist of a tiered storage taxonomy that distinguishes between datasets based on their sensitivity and correspondingly restricts access to different research groups. Interpreting tiered storage and access as two sides of the same coin, we reference existing models that are based on dataset risk levels and propose a framework for the NRC that aims to achieve the dual goals of streamlining the process of enabling research access to government data while maintaining privacy and security.

FedRAMP: A tiered framework for data storage on the cloud

One type of tiered storage taxonomy already exists for third party government cloud services in one of the federal government's major cybersecurity frameworks, the Federal Risk and Authorization Management Program (FedRAMP).[33] Enacted in 2011, the framework was designed to govern all federal agency cloud deployments, with certain exceptions detailed in Chapter 8 of this White Paper. FedRAMP offers two paths for cloud services providers to receive federal authorization. First, an individual agency may issue what is known as an authority-to-operate (ATO) to a cloud service provider after the provider's security authorization package has been reviewed by the agency's staff and the agency has identified any shortcomings that need to be addressed.[34] These types of ATOs are valid for each vendor across multiple agencies, as other agencies are permitted to reuse an initial agency's security package in granting ATOs. The second option available to cloud services providers is to obtain a provisional ATO from the FedRAMP Joint Authorization Board, which consists of representatives from the Department of Defense (DOD), the Department of Homeland Security (DHS), and the

General Services Administration (GSA). These provisional ATOs offer assurances to agencies that DHS, DOD, and the GSA have reviewed security considerations, but before any specific agency is allowed to use a vendor's services, that agency must issue its own ATO.[28] In both the first and second cases, FedRAMP categorizes systems into low, moderate, or high impact levels (see Table 2).

Because FedRAMP requirements apply to all federal agencies when federal data is collected, maintained, processed, disseminated, or disposed of on the cloud, the NRC itself will need to be compliant with FedRAMP security standards irrespective of the organizational form it takes.[35] Every dataset brought on to the NRC would need to be reviewed under FedRAMP with appropriate access levels. If a cloud service has already been evaluated under FedRAMP because it was used in the past to house federal data, the service can inherit the same FedRAMP compliance level in the NRC without an additional evaluation.[36]

Besides classifying datasets, the other function of FedRAMP is to identify a comprehensive set of "controls," i.e., requirements and mechanisms that the cloud service providers must implement before the government dataset can be housed on them.[37] They are based on the National Institute of Standards and Technology (NIST) Special Publication 800-53, which provides standards and security requirements for information systems used by the federal government.[38]

These controls range widely and include requirements such as ensuring that the organization requesting certification "automatically disables inactive accounts," "establishes and administers privileged user accounts in accordance with a role-based access scheme that organizes system access and privileges into roles," "provides security awareness training on recognizing and reporting potential indicators of insider threat," or develops regular security plans in the event of a breach.[39] Requirements get more strenuous for FedRAMP "high impact" data (e.g., creating system level air-gaps to protect sensitive data).[40]

LEVEL	TYPE OF DATA	IMPACT OF DATA BREACH	NUMBER OF CONTROLS
Low-impact risk - Low baseline - Low-impact SaaS	Data intended for public use	Limited adverse effects; preserves the safety, finances, reputation, or mission of an agency	125
Moderate-impact risk - E.g., personally identifiable information	Controlled unclassified data not available to the public	Can damage an agency's operations	325
High-impact risk - E.g., law enforcement, healthcare, emergency services	Sensitive federal information	Catastrophic impacts such as shutting down an agency's operations, causing financial ruin, or threatening property or life	421

Table 2: FedRAMP levels are designated based on the degree of risk associated with the breach of an information system. The security baseline levels are based on confidentiality, availability, and integrity, as defined in Federal Information Processing Standard 199.[41]

There can be significant costs with obtaining these certifications and creating compliance plans, even if the underlying technical specifications can be addressed or already exist. A key issue for structuring the NRC is that the principal burdens of ensuring FedRAMP compliance should fall with NRC institutional staff, not originating agencies or individual academic researchers. As part of the FedRAMP certification process, NRC staff will have to consider how to give access to PIs in compliance with FedRAMP rules, but that process can and should avoid requiring originating agencies or individual universities to incur substantial expenses associated with hiring consultants and attorneys to certify FedRAMP compliance.[42]

While FedRAMP sets out common standards for cloud storage of government data within agencies,[43] it is an exception to an otherwise balkanized federal data-sharing standards landscape,[44] though it does not facilitate data exchange. The NRC needs to maintain compliance with not only FedRAMP requirements but also the requirements of any agency it is partnering with for data access.[45] Advocates interested in increasing government data availability have long fought to establish a universal FedRAMP equivalent across different agencies that provides shared standards for data sharing based on data sensitivity.[46] As we discuss in Chapter 8, establishing such universal, "centralized" security standards not only ensures internal uniformity but also removes barriers to data sharing.

The NRC's implementation of FedRAMP standards can also provide partnering agencies an important opportunity to reexamine their own standards and share best practices with one another.[47] This could involve raising or lowering requirements that are out of date,[48] given the current threat to the environment and research needs. The NRC can take inspiration from agencies' best practices, as well as from FedRAMP to develop a common NRC standard for determining data to be high, moderate, or low risk, as well as what consequences should flow from that assessment. In the later section on strategic considerations, we discuss how to enable this process by incentivizing agencies to participate in the NRC and selecting datasets that present a lower privacy and security risk.

In addition, given the diversity of data types and sources that could be stored on the platform, NRC policy should ensure that standards and protections exist for data storage in areas where FedRAMP has blind spots. FedRAMP is in part animated by risks from malicious actors like cybercriminals or adversarial foreign governments, but as we discuss in Chapter 6, privacy risks may arise even for the intended use case of analysis by NRC researchers. Of particular concern are instances where disparate datasets are combined, which may allow new inferences that make previously anonymous data individually identifiable, even when the data itself did not contain identifiable information.[49] Such combinations may also alter the original risk level of the data, creating an output that merits a higher risk classification. Furthermore, machine-learning models and representations may unintentionally reveal properties of the data used to train them,[50] and dissemination of these models could pose privacy risks.

This is not a challenge unique to the NRC; the U.S. Census Bureau and other government agencies engaged in data linkage have also had to develop means to address this issue.[51] One solution involves applying methods of additional noise to the data (differential privacy) in order to obfuscate individual data while preserving the data's utility for research. We discuss it and other privacy-enhancing technologies in greater detail in Chapter 7.[52] However, privacy-enhancing technologies are no panacea, and depending on the nature of the particular dataset, the goals of ensuring anonymization, while also enabling researchers to access fine-grained data can conflict.

The NRC can also draw from the "Five Safes" data security framework used by the UK Data Service,[53] the Federal Statistical Research Data Centers Network, and the Coleridge Initiative, a model centered on data, projects, people, access settings, and outputs.[54] The implementation of the 2019 Evidence Act is already using a similar Five Safes framework in making determinations around data linkage.[55] Through a combined framework, the NRC could place different anonymization requirements on datasets, depending on the circumstances of their access and the privacy

agreements through which they were collected. Similarly, the NRC could control the dissemination scope of models, code, and data, depending on the sensitivity. Theoretical identifiability is less likely to be a concern when access and dissemination is restricted and the data is of a less sensitive nature or is not about individuals at all.[56]

Facilitating Researcher Access with a Tiered Access Model

How should researchers gain access to specific data resources? Currently, approval proceeds on an agency-by-agency basis.[57] Just as the value of the NRC for supporting AI research will depend in part on the extent to which it can bring together datasets from different agencies, it will also depend on the extent to which it can streamline the process for accessing data. One way to achieve this streamlining will be through a tiered access system for the NRC users, similar to FedRAMP's tiered system for storing federal data on the cloud, where higher tiers would enable access to higher-risk data, subject to the other requirements on compute and data use. We discuss this access system in more depth in Chapter 7.

Chapter 2 made the case that compute access should start with PIs at academic institutions. This authorization can also serve as the baseline, where all NRC-registered PIs can freely access and use low-risk datasets on the NRC. Additional tiers would impose more requirements, such as citizenship, security clearance, distribution restrictions, or compute and system restrictions. These access tiers will be similar to those used for determining FedRAMP classification for data storage, but while access and storage sensitivity may invoke similar considerations; they might not necessarily be the same.

CASE STUDY: COLERIDGE INITIATIVE (ADMINISTRATIVE DATA RESEARCH FACILITY)

In partnership with the Census Bureau and funding from the Office of Management and Budget, the Coleridge Initiative, a nonprofit organization, launched the Administrative Data Research Facility (ADRF), a secure computing platform for governmental agencies to share and work with agency micro-data.[58] The ADRF is available on the Federal Risk and Authorization Management Program (FedRAMP) Marketplace and has a FedRAMP Moderate certification. Currently, the platform supports over 100 datasets from 50 agencies.[59]

The ADRF provides access to agency-sponsored researchers and agency-affiliated researchers going through the ADRF training programs for free. Over the past three years, over 500 employees from approximately 100 agencies have gone through ADRF training programs.[60]

The ADRF provides a shared workspace for projects and the Data Explorer, a tool to view an overview and metadata (name, field description, and data type) of available datasets on the ADRF.[61] In order to access restricted data, users must meet review requirements set by the agency providing the data. In order to export data, users must go through a unique "Export Review" process.[62] The ADRF has a highly involved default review process, requiring researchers to submit all code and output for the project for approval to the data steward and generating additional charges, if requesting export of more than 10 files.[63] The agency providing the data can also amend the default review process, if it wishes to do so.

Prior to transferring data files, the ADRF provides an application for data hashing to safely transmit data.[64] The ADRF also follows the "Five Safes" security model used by other government agencies, such as the UK Data Service.[65]

Data stewardship for the ADRF is defined in compliance with the Title III of the Evidence-Based Policymaking Act of 2018.[66] Once a restricted dataset is shared with the ADRF, one person within the agency will be assigned the data steward for all project requests. From there, procedures are developed with the agency, in terms of expectations for how the data will be protected, authorized users, and audit procedures for continued compliance.

Data stewards have access to an online portal in the ADRF. All project requests for specific data are routed to the data steward through this proposal. Once access has been granted, the data steward also has options to monitor the project for compliance.

KEY TAKEAWAYS

- ***Balances providing consistent restricted data access with agency requirements:*** The ADRF balances building in each restricted data set's access and export review form in a consistent manner into the data portal with agency requirements for data access and export. This allows agencies to control access to their data, while providing a single point of entry for researchers. Currently, the platform only supports data consistent with a FedRAMP Moderate certificate.
- ***Standardized for both users and data stewards:*** Along with each data access for users, a point of contact at the agency providing the data is given access to the platform as well. This allows easy access to approve and track projects, and work with the ADRF on access requirements.
- ***Five Safes' data security framework:*** The ADRF ensures data security by focusing on five aspects—data, projects, people, settings of access, and outputs.
- ***Training as a core function:*** The ADRF hosts workshops and trains government employees and other researchers on data use.

Existing models for researcher access to sensitive datasets can help paint a picture of how the NRC might maintain and monitor a tiered access system. The NRC can emulate both the Coleridge Initiative and Stanford's Center for Population Health Sciences (PHS),[67] for instance, which serve as data intermediaries, in facilitating access to government data. Indeed, these intermediaries have been documented as effective means to overcome barriers to data-sharing because they, at their core, negotiate and streamline relationships between data contributors and users.[68] For example, as a trusted intermediary, the NRC could centralize the DUA intake process by promulgating a universal standard form for agency DUAs.[69]

Furthermore, similar to the Coleridge Initiative example, a designated representative(s) within the agency could be assigned as the data steward for all project requests for a certain restricted dataset. Any project requiring access to data in higher tiers could commence only after its proposal was reviewed and approved by a relevant representative. Because NRC access begins with PIs, researchers would also have to obtain approval from their university Institutional Review Boards (IRBs), as needed. After project approval and NRC researcher clearance, data would be made available through the NRC's secure portal. Any violations of the terms of use or subject privacy could result in penalties ranging from a demotion of access tier to removal of NRC privileges or professional, civil, or criminal penalties, as relevant.

The NRC can emulate both the Coleridge Initiative and Stanford's Center for Population Health Sciences (PHS), for instance, which serve as data intermediaries, in facilitating access to government data.

CASE STUDY: STANFORD CENTER FOR POPULATION HEALTH SCIENCES

The Stanford Center for Population Health Sciences (PHS) provides a growing set of population health-related datasets and access methods to Stanford researchers and affiliates.[70] The PHS Data Ecosystem hosts high-value datasets, data linkages and filters, and analytical tools to aid researchers. The PHS partners with a wide range of public, nonprofit, and private entities to license population-level datasets for university researchers, ranging from low-risk, public datasets to restricted data containing Protected Health Information (PHI) and Personally-Identifiable Information (PII), such as Medicare, commercial claims such as Optum and Marketscan data,[71] and electronic medical records.

KEY TAKEAWAYS

- ***Mixed data architecture, yet consistent user experience:*** PHS utilizes a mix of on-premises and cloud data services, but still seeks to provide a consistent user experience.
- ***Restricted data access has a single point of entry:*** The PHS Data Portal standardizes, centralizes, and simplifies data access requirements and trainings, rather than pointing users to a time-consuming process working with each data steward directly.
- ***Reduces costs and time associated with procuring data:*** PHS leverages existing relationships with agencies to consolidate datasets in a single portal, saving researchers the time and money necessary to gain access through individual agency requests.

CASE STUDY: STANFORD CENTER FOR POPULATION HEALTH SCIENCES (CONT'D)

In addition to secure data storage and computational tools for researchers, PHS provides standardized and well-documented data access and management protocols, which increases data proprietor comfort with sharing data. PHS also has full-time staff who cultivate and maintain relationships with organizations holding data. This allows PHS to work with these groups to centralize data hosting and provide secure access to a wide array of researchers.

The PHS Data Portal is hosted on a third-party platform that enables data discovery, exploration, and clearly delineated, standardized steps for data access. The third-party platform, Redivis, utilizes a four-tier access system: (1) overview of data and basic documentation; (2) metadata access, including definitions, descriptions, and characteristics; (3) a 1 or 5 percent sample of the dataset; and (4) full data access.[72]

If data is classified as public, researchers can access it using specialized software, or simply download it directly.[73] For restricted data, the portal has forms integrated to easily apply for access.[74] After identifying the dataset, the researcher must apply for membership in the organization hosting the data.[75] An administrator of the organization owning the dataset can set member and study requirements that must be met, including training and institutional qualification, in order to access the data. Member applications can be set to auto-approval or require administrative approval. Once access has been granted to a data set, researchers can manipulate the data using specialized software. Usage restrictions are also specified individually on each dataset to control whether full, partial, or no output can be exported, and what review level is required for exporting. All applications for data and export are handled directly on the Data Explorer platform.

Currently, the PHS Data Portal is primarily for Stanford faculty, staff, students, or other affiliates.[76] Even with affiliate status, certain commercial datasets may require further data rider agreements for access. Non-Stanford collaborators must complete all of the same access requirements as Stanford affiliates, plus any requirements imposed by their own institution. Additionally, a "data rider" agreement on the original DUA is frequently necessary.[77]

To work with restricted data, the PHS provides two computing services for high-risk data: (1) Nero, with both an on-premises and Google Cloud Platform (GCP) platform versions; and (2) PHS-Windows Server cluster.[78] Both are managed by the Stanford Research Computing Center (SRCC). Both services are HIPAA compliant.[79] Unrestricted data can be used on any of Stanford's other computational environments (Sherlock, Oak) or simply downloaded to the researcher's local machine.

PROMOTING INTERAGENCY HARMONIZATION AND ADOPTION OF MODERN DATA ACCESS STANDARDS

The federal data-sharing landscape suffers from divergent standards and practices, and individual agencies, left alone, have traditionally faced high hurdles to harmonizing and modernizing their data access standards.[80] As we have discussed, this state of affairs presents formidable barriers to AI R&D from a researcher perspective, but is also problematic both from an agency and societal perspective. As a report by the Administrative Conference of the United States finds from surveying the use of AI in the federal government, nearly half of agencies have experimented with AI to improve decision-making and operational capabilities, but they often lack the technical infrastructure and data capacity to use modern AI techniques and tools.[81] The lack of a modern, uniform standard for data sharing in AI research, therefore, makes it harder for agencies to realize gains in accuracy, efficiency, and accountability, which subsequently impacts citizens downstream, who are affected by agency decisions.[82]

The lack of a modern, uniform standard for data sharing in AI research makes it harder for agencies to realize gains in accuracy, efficiency, and accountability, which subsequently impacts citizens downstream, who are affected by agency decisions.

The NRC can help overcome agency reluctance to share data by enabling access to agencies to compute on their own data. This would solve at least two crucial problems for government agencies. First, access to the NRC's collective computing resources would overcome some difficulties that agencies have traditionally faced in setting up their own compute resources.[83] Second, facilitating agency access to modern data and compute resources would attract and build further in-house government AI expertise.[84] From a societal perspective, this could increase the government's capabilities in the responsible adoption of AI, help reduce the cost of core governance functions, and increase agency efficiency, effectiveness, and accountability.[85]

The NRC can also learn from and align with other initiatives to harmonize and modernize standards. The Evidence Act—which requires agencies to appoint chief data and evaluation officers—is one example. The legislation authorizing the creation of the NRC could provide a federal mandate to encourage adoption of sharing best practices.[86] However, as we discuss in Chapter 5, a federal mandate alone, without any additional aid or incentives, may not be enough to incentivize harmonization of data access and sharing standards.[87] The Task Force should therefore consider bundling the mandate with additional benefits, such as providing funding to assist agencies in expanding their technical or staff capabilities in furtherance of the NRC and the national AI strategy. The NRC is aligned with the existing bipartisan case for the National Secure Data Service (NSDS) (described in the case study below), a service that would facilitate researcher access to data with enhanced privacy and transparency, recommended by the Commission on Evidence-Based Policymaking in 2018. Both the NRC and NSDS are complementary data-sharing initiatives that have the potential to considerably improve public service operational effectiveness. We elaborate further on the NSDS proposal in Chapter 5. Lastly, training programs are promising avenues to increase NRC adoption and agency support. For example, as described in the case study above, the Coleridge Initiative has hosted workshops to train over 500 employees from approximately 100 agencies on data use over the past three years.

CASE STUDY: THE EVIDENCE ACT

In pursuit of greater, more secure access to and linkage of government administrative data, a bipartisan Commission on Evidence-Based Policymaking was set up by Congress in March 2016. The commission's final report[88] included 22 recommendations for the federal government to build infrastructure, privacy-protecting mechanisms,[89] and institutional capacity to provide secure access to public data for statistical and research purposes. One recommendation was to create a "National Secure Data Service" (NSDS) to facilitate access to data for the purpose of building evidence, while maintaining privacy and transparency. Through this service, the NSDS could help researchers by temporarily linking existing data and providing secure access, without itself creating a data clearinghouse.

The Foundations for Evidence-Based Policymaking Act of 2018[90] created some of the legislative footing for the commission's recommendations. In particular, it created new roles for chief data, evaluation, and statistical officials, and sought to increase access and linkage of datasets previously within the scope of the Confidential Information Protection and Statistical Efficiency Act (CIPSEA).[91]

Finally, the 2020 Federal Data Strategy and associated Action Plan[92] sought to put those legislative provisions into action. The strategy included plans to improve data governance, to make data more accessible, to improve government use of data, and to boost the use and quality of data inventories, metadata, and data sensitivity.

The central remaining step envisioned by the initial Evidence-Based Policymaking Commission is a National Secure Data Service (NSDS) modeled on the UK's Data Service.[93] The UK's Data Service provides access to a range of public surveys, longitudinal studies, UK census data, international aggregate data, business data, and qualitative data. Alongside access, it provides guidance and training for data use, develops best practices and standards for privacy, and has specialized staff who apply statistical control techniques to provide access to data that are too detailed, sensitive, or confidential to be made available under standard licenses.

KEY TAKEAWAYS

- ***Priority data sharing where there is a public service operational case:*** There is precedent in large-scale administrative data-sharing initiatives justified on grounds of improvements to public service operational efficiency and effectiveness.
- ***National Secure Data Service (NSDS) initiative:*** The NSDS is bolstered by bipartisan support.
- ***Institutional models that balance external, innovative talent and internal, cross-agency influence:*** A Federally Funded Research and Development Center (FFRDC) model, housed within an existing agency (NSF), may balance the ability to bring in external talent with internal agency influence.

SEQUENCING INVESTMENT INTO DATA ASSETS

Given the significant hurdles in negotiating data access, the NRC will need to strategically sequence which agencies and datasets to focus on for researcher use. The federal government collects petabytes of data,[94] each with varying degrees of restrictions or openness. In considering which datasets to prioritize, the NRC can draw from the example of other data-sharing initiatives, as well as focus on data sets in the short term that do not pose complex challenges with regards to data privacy or sharing. One private sector example is Google Earth Engine, which aggregated petabytes (approximately 1 million gigabytes) of satellite images and geospatial datasets, and then linked that access to Google's cloud-computing services to allow scientists to answer a variety of crucial research questions.[95] This process of aggregating complex data and hosting it in a friendly computing infrastructure to facilitate research, demonstrates the compelling value of coupling compute and data. As another example, ADR UK identifies specific areas of research that are of pressing policy interest, such as "world of work,"[96] and prioritizes data access for researchers working on those topics. The UK Data Service offers datasets derived from survey, administrative and transaction sources, including productivity data from the Annual Respondents Database,[97] innovation data from the UK Innovation Survey,[98] geospatial data from the Labour Force Survey,[99] Understanding Society,[100] and sensitive data about childhood development.[101]

When prioritizing datasets and agencies for NRC partnership, we recommend the following criteria:

- *Data that is valuable to AI researchers, but is not currently available in a convenient form.* For example, in a July 2019 request for comments, the Office of Management and Budget (OMB) asked members of the public to provide input on characteristics of models that make them well-suited to AI R&D, what data is currently restricted, and how liberation of such data would accelerate high-quality AI R&D.[102] In one response, the Data Coalition argued that controlled release of private but structured indexed data in data.gov would be valuable for research.[103] The Data Coalition also urged agencies to consider releasing raw, unstructured datasets, such as agency call center logs, consumer inquiries and complaints, as well as regulatory inspection and investigative reports.[104] Another example of data that is currently challenging to access, but is a matter of public record, is electronic court records housed in a system by the Administrative Office of the U.S. Courts.[105]

- *Data housed within agencies that have statutory authority to share data and/or that have previous data-sharing experience*. The Census Bureau, for instance, has greater existing statutory interagency linkage than other agencies, and has preexisting substantial in-house data analysis expertise.[106] The U.S. Bureau of Labor Statistics has an existing process for sharing restricted datasets (in the categories of employment and unemployment, compensation and working conditions, and prices and living conditions) with researchers.[107]

- *Data with limited privacy implications.* For example, agencies whose data concerns natural phenomena, rather than individuals, may be easier to manage from a privacy perspective—e.g., NASA, the US Geological Service, and the National Oceanic and Atmospheric Administration. Datasets like those housed in NASA's Planetary Data System,[108] but that are not easily available to researchers, may serve as a valuable starting point for the NRC. Increasing the availability and interoperability of datasets from these agencies would advance the core mission of the NRC and could be done without jeopardizing individual privacy.

Chapter 4: Organizational Design

What institutional form should the NRC take? Two overarching considerations are: (1) ease of access to data; and (2) ease of coordination with compute resources.[1] As we discussed in Chapter 3 and will detail in depth in Chapter 5, the federal data-sharing landscape among agencies is highly fragmented, with many agencies reluctant to or legally constrained from sharing their data. The NRC will need to coordinate between the entities supplying compute infrastructure and researchers themselves. As the NRC's goal is to provide researchers with access to government data *and* high-performance computing power, one without the other will fall short of achieving the NRC's mission.

Drawing on extensive work in support of the Evidence Act, we recommend the use of Federally Funded Research and Development Centers (FFRDCs) and private-public partnerships (PPPs) as possible organizational forms for the NRC. We recommend the creation of an FFRDC at affiliated government agencies in the short term, as we believe this path allows for the easiest facilitation of both the compute infrastructure and access to government data. In the longer term, the establishment of a PPP could facilitate greater data sharing and access between the public and private sectors. Importantly, other options include creating an entirely new federal agency or bureau within an existing agency. While these options might simplify coordination with compute resources, both pose challenges, with respect to data accessibility and interagency data sharing.

KEY TAKEAWAYS

- In the short term, the NRC should be instituted as a Federally Funded Research and Development Center (FFRDC), which would reduce the significant costs of securing data from federal agencies.
- In the longer term, a well-designed, public-private partnership (PPP), governed by officers from Affiliated Government Agencies, academic researchers, and representatives from the technology sector, could increase the quantity and quality of R&D, and reduce maintenance costs.
- Instituting the NRC as a standalone federal agency or bureau would face numerous challenges, notably in securing access to data housed in other agencies.

FEDERALLY FUNDED RESEARCH AND DEVELOPMENT CENTER

FFRDCs are quasi-governmental nonprofit corporations sponsored by a federal agency but operated by contractors, including universities, other nonprofit organizations, and private-sector firms.[2] The FFRDC model confers the benefits of a close agency relationship, alongside independent administration, in facilitating access to data. Due to their intimate subcontracting relationships with their parent agency, all FFRDCs benefit from data access that goes "beyond that which is common to the normal contractual relationship, to Government and supplier data, including sensitive and proprietary data."[3]

A recent report by Hart and Potok on the National Secure Data Service (NSDS) (see case study in Chapter 3) also supports the FFRDC model as an optimal way to facilitate access to and linkage of government administrative data.[4] The report considered FFRDCs, alongside such other institutional forms, as creating an entirely new agency, housing the NSDS in an existing agency, and developing a university-

led, data-sharing service, but the report ultimately recommended the FFRDC model for several reasons. An FFRDC can scale quickly, because it can access government data and high-quality talent more easily than other options.[5] An FFRDC can also leverage existing government expertise. The NSF, for instance, already sponsors five separate FFRDCs and has extensive experience cultivating and maintaining networks of researchers.[6]

However, the FFRDC model comes with a few limitations. First, an FFRDC's role is restricted to research and development for their sponsoring agency that "is closely associated with the performance of inherently governmental functions."[7] Thus, it would be important to ensure alignment during the contracting phase with the NRC's core functions.

Second, the success of an FFRDC model for the NRC will depend on the ability of the sponsoring agency to gain cooperation across the federal government to provide data needed for research. One way to do this would be for multiple agencies to co-sponsor the FFRDC, reducing contracting friction for datasets.[8] Another option would be to create *multiple* FFRDCs housed in different agencies, incentivizing each of those agencies to share their data with the respective FFRDC. An analogous example could include the National Labs as a network, where each National Lab would be an *instantiation* of the NRC within its own relevant agency.[9]

Third, multiple FFRDCs would require separate processes for compute resources. In the short term, the NRC may alleviate this problem by contracting for commercial cloud credits, which is likely already the short-term solution for the NRC to provide compute access. As discussed earlier, private sector cloud providers already have extensive experience in providing compute resources to the government[10] and to academic institutions.[11] Familiarity with these private cloud providers may reduce the friction in allocating compute among researchers at multiple FFRDCs.

In the longer term, the FFRDC model may not be the most efficient. From a cost and sustainability perspective, FFRDCs have traditionally suffered from significant overruns, as they "operate under an inadequate, inconsistent patchwork of federal cost, accounting and auditing controls, whose deficiencies have contributed to the wasteful or inappropriate use of millions of federal dollars."[12] Another concern is that, historically, FFRDC infrastructure has not been routinely updated. A 2017 Department of Energy report highlighted that FFRDC infrastructure was inadequate to meet the mission.[13] NASA's Inspector General also highlighted that more than 50 percent of the Jet Propulsion Laboratory (a NASA FFRDC) equipment was at least 50 years old.[14] If an FFRDC version of the NRC experiences these same challenges, we recommend that the NRC, in the long run, switch to a public-private partnership model.

CASE STUDY: SCIENCE & TECHNOLOGY POLICY INSTITUTE (STPI)

STPI is an FFRDC chartered by Congress in 1991 to provide rigorous objective advice and analysis to the Office of Science and Technology Policy and other executive branch agencies.[15] STPI is managed by the Institute for Defense Analyses (IDA), a nonprofit organization that also manages two other FFRDCs: the Systems and Analyses Center and the Center for Communications and Computing.[16] IDA has no other lines of business outside the FFRDC framework.[17]

STPI's primary federal sponsor is the National Science Foundation, but research at STPI is also co-sponsored by other federal agencies, including the National Institute of Health (NIH), Department of Energy (DOE), Department of Transportation (DOT), Department of Defense (DOD), and Department of Health and Human Services (HHS).[18] Due to the "unique relationship" between an FFRDC and its sponsors, STPI "enjoys unusual access to highly classified and sensitive government and corporate proprietary information."[19]

NSF appropriations provide the majority of funding for STPI, including $4.7 million in FY 2020,[20] but a limited amount of funding is also provided from other federal agencies.[21] STPI has approximately 40 full-time employees and has access to the expertise of IDA's approximately 800 other employees.[22] As an FFRDC, STPI may also contract for expertise, as required for a particular project.[23] The statute specifying STPI's duties also directs it to consult widely with representatives from private industry, academia, and nonprofit institutions, and to incorporate those views in STPI's work to the maximum extent practicable.[24]

STPI is also required to submit an annual report to the president on its activities, in accordance with requirements prescribed by the president,[25] which provides additional accountability for the FFRDC. According to STPI's 2020 report, STPI worked across multiple federal agencies, supporting them on 48 separate technology policy analyses throughout 2020.[26]

KEY TAKEAWAYS

- ***Multiple agency co-sponsors:*** While STPI's primary sponsor is the National Science Foundation, a number of other agencies also co-sponsor STPI, reducing difficulties in accessing data across agencies.
- ***Expertise:*** While STPI is staffed with its own employees, it can also tap into expertise from the hundreds of employees at the Institute for Defense Analyses (IDA), the organization that manages STPI. As an FFRDC, STPI can also contract for additional expertise as required.

A PUBLIC-PRIVATE PARTNERSHIP (PPP)

A Public-Private Partnership (PPP) would create a partnership between federal agencies and private-sector organizations to jointly house and manage data-sharing efforts and run compute infrastructure. Because different agencies and private-sector members may have different contracting preferences, intellectual property goals, and security allowances for data access, creating a data-sharing partnership within this patchwork framework could be challenging in the immediate future. Nonetheless, PPPs can provide a number of long-term benefits, as they have been used successfully as data clearinghouses to produce, analyze, and share data between the public and private sector.[27] Indeed, recognizing the benefits of the PPP model, the European Union has launched a new initiative called the Public Private Partnerships for Big Data that will offer a secure environment for cross-sector collaboration and experimentation using both commercial and public data.[28] In general, PPPs for data-sharing can increase the quality and quantity of R&D, increase the value and efficiency of sharing public sector data, and reduce the long-run cost necessary to manage and maintain the data-sharing infrastructure.[29]

CASE STUDY: ALBERTA DATA PARTNERSHIPS (ADP)

Founded in 1997, the ADP PPP is designed to provide long-term management of comprehensive digital data sets for the Alberta market.[30] The PPP is structured as a joint venture between ADP, a nonprofit, and Altalis Ltd. whereby the ADP is the "custodian" of government data and Altalis is the "operator."[31] More specifically, geospatial data is owned by the provincial government, but exclusive licensing arrangements are granted to ADP to allow for sales.[32] Meanwhile, Altalis, under the direction and oversight of ADP, builds software to securely load and distribute these provincial spatial datasets to users. Altalis also provides training to end users and is responsible for cleaning, updating, and standardizing datasets.[33]

In choosing its "operating partner" (i.e., Altalis) for the joint venture, the ADP board initially issued a "Request for Information" that solicited proposals from private-sector companies whose core business was the improvement, maintenance, management, and distribution of spatial data.[34] The ADP board ultimately chose Altalis, not only because it had the superior offering and existing capabilities, but also because Altalis was willing to take on all the investment required, at its own risk, to build and operate the ADP system in accordance with ADP specifications.[35]

Today, all Altalis and ADP costs are covered by the operations of the joint venture.[36] The joint venture earns revenues through, for instance, through directed project funding and data access fees from stakeholders, which include municipalities, regulatory agencies, energy, forestry, and mining organizations.[37] Any profits from the joint venture are split roughly 80/20 between Altalis and ADP, respectively, and ADP subsequently uses its profit share to reinvest in data and system improvements.[38]

The ADP PPP claims to have generated efficiencies for data sharing. For instance, the ADP estimates that a traditional government-only approach to maintaining and distributing datasets would have ranged between $65 million and $120 million cumulatively since ADP's inception, and ADP claims to have provided its users with $6.8 million in cost savings.[39]

KEY TAKEAWAYS

- ***Utilizing joint ventures:*** The ADP PPP uses a joint venture agreement to set the terms and conditions for the partnership, whereby one partner is the custodian for government data, and the other is the operator that builds and maintains the software that facilitates data-sharing.
- ***Revenue-sharing agreements:*** A shared revenue model assures contributions from, and realization of value to, each stakeholder. The ADP subsequently reinvests its profits into improving the system.
- ***Significant efficiencies:*** According to the ADP, there are lower costs to creating and maintaining the ADP than under a conventional approach.

A PPP model could reduce the friction of coordination between data and compute. One example of using a PPP for compute resources is the COVID-19 High-Performance Computing Consortium, spearheaded by the Office of Science and Technology Policy, DOE, NSF, and IBM.[40] Drawing on the experience of XSEDE, the consortium has 43 members from the public and private sectors that volunteer free compute resources to researchers with COVID-19-related research proposals.[41] The voluntary nature of compute provisioning, in this instance, provides benefits to both the researchers, who gain immediate access to compute, and the consortium members, who contribute to innovation and reap public relations benefits.

We also acknowledge that the evidence around the efficacy of PPPs is contested.[42] Indeed, there is no one-size-fits-all PPP model; PPPs differ vastly, according to the responsibilities allocated between the private sector and the public sector, and the success of a PPP can depend on its structure.[43] According to a RAND Report of 30 case studies of successful public-private data clearinghouses, these clearinghouses have widely different organizations, access requirements, and strategies for managing data quality.[44] Such decision points are crucial. For example, some scholars emphasize the need for a trusted environment for the private and public sectors to handle privacy and ethics violations in sensitive industries.[45] Similarly, in the siloed federal data-sharing context, a PPP must consider how to divide functions in tackling these additional considerations in privacy, ethics, security, and intellectual property.

THE NRC AS A GOVERNMENT AGENCY

The NRC could also be constructed as a new government agency or bureau. The main advantages to this model would be the development of a distinct public-sector institution, devoted to AI compute and data. The NRC could be to cloud and data what the U.S. Digital Service is to government information technology. Such an agency would have to be established by statute or executive mandate. Enabling legislation could create dedicated, professional staff to build and develop the NRC, vest the NRC with authority to mandate interagency data sharing, and create a long-term plan that is informed by the National AI Strategy.

There are, however, significant disadvantages to creating a new agency or bureau. First, the NRC could lay claim to no government datasets at all, and could subsequently encounter significant headwinds with having to negotiate with each originating agency for data, not to mention the constraints under the Privacy Act, discussed in Chapter 5. That said, enabling legislation could exempt the agency from the Privacy Act's data linkage prohibitions and transfer litigation risk for data leakages to the new agency. Second, a new agency may face greater challenges in recruiting top-flight talent.[46] According to the 2020 Survey on the Future of Government Service, a majority of respondents at federal agencies agreed that they often lose good candidates because of the time it takes to hire, and less than half agreed that their agencies have enough employees to do a quality job.[47] Moreover, many respondents highlighted inadequate career growth opportunities, inability to compete with private-sector salaries, and lack of a proactive recruiting strategy as major factors contributing to an inadequately skilled workforce in federal agencies.[48] FFRDCs, in contrast, can be negotiated with existing organizations, making the startup costs potentially lower. Third, while national laboratories have expertise contracting with entities to construct high-performance computing facilities, it is unclear how a new federal agency/office would approach such a task. It is one thing for an entity like the U.S. Digital Service to help develop IT platforms for U.S. agencies; it is another to simultaneously build a very large supercomputing facility and solve longstanding challenges with data access. Finally, it will be important to isolate the research mission of the NRC from political influence. To the extent that a new agency might provide less isolation from changes in presidential administrations and politically appointed administrators, this is an important consideration.

While these disadvantages are considerable, ambitious legislative action could, in fact, make a new government agency a viable option.

Chapter 5: Data Privacy Compliance

The vision motivating the NRC is to support academic research in AI by opening access to both compute and data resources. Federal data can fuel basic AI research discoveries and reorient efforts from commercial domains toward public and social ones. As stated in the NRC's original call, "Researchers could work with agencies to develop and test new methods of preserving data confidentiality and privacy, while government data will provide the fuel for breakthroughs from healthcare to education to sustainability."[1]

But is an NRC seeded with public sector data, particularly administrative data from U.S. government agencies, even possible given the legal constraints? Research proposals that sweep broadly across agencies for personally identifiable or otherwise sensitive data[2] will rightly trigger concerns about potential privacy risk. The Privacy Act of 1974, the chief federal law governing data collected by government agencies, fundamentally challenges the notion of an NRC as a one-stop shop for federal data. Its research exceptions leave some uncertainty about open-ended research endeavors that go beyond statistical research or policy evaluation supporting an agency's core mission. Even if agencies deemed such research possible, researchers would be subject to access constraints and the data itself may potentially require technical privacy treatments.

We make the following recommendations regarding data privacy and the NRC. First, agencies may be able to share anonymized administrative data with the NRC within the boundaries of the Privacy Act for the purposes of AI research, based on the Act's statistical research exemptions. Second, the NRC will require a staff of privacy professionals that include roles tasked with legal compliance, oversight, and technical expertise. These professionals should build relationships with peers across agencies to facilitate data access. Third, the NRC should explore the design of virtual "data safe rooms" that enable researchers to access raw administrative microdata in a secure, monitored, and cloud-based environment. Fourth, we recommend the NRC Task Force engage the policy and statistical research communities, and consider coordination with proposals for a National Secure Data Service, which has grappled extensively with these issues.

This chapter proceeds as follows. We first review the existing laws that apply to government agencies and the restrictions they impose on data access and sharing. We then describe current agency practices for sharing data with researchers and agencies under the Privacy Act. Last, we assess the implications of current legal constraints on NRC data sharing and the most important cognate proposal to promote data sharing under the Evidence Act.

KEY TAKEAWAYS

- Agencies may share anonymized administrative data with the NRC under the statistical research exemption of the Privacy Act.
- An agency's willingness and ability to share data may depend on the extent to which a proposed research project aligns with an agency's core purpose.
- The NRC will require a staff of privacy professionals for legal compliance, oversight, and technical expertise.
- Individually identifiable or sensitive data will face obstacles to release and may warrant technical privacy and/or tiered access measures.

We note at the outset that this chapter largely takes existing statutory constraints as a given. At a macro level, however, the challenges in data sharing also suggest that an ambitious legislative intervention could overcome many existing constraints, such as by statutorily (a) exempting the NRC from the Privacy Act's prohibition on data linkage; (b) granting the NRC the power to assume agency liabilities for data breaches; (c) mandating that agencies transfer any data that has been shared under a data use agreement or Freedom of Information Act (FOIA) request to the NRC; and (d) requiring IT modernization plans to include provisions for data-sharing plans with the NRC.[3]

THE PRIVACY ACT

Data privacy issues are at the core of debates about sharing data, and the NRC will be no exception. Most data privacy debates in the U.S. today focus on the consumer data sector where data protection laws in the U.S. are limited to nonexistent. In contrast, many U.S. government agencies are subject to a robust privacy law, the Privacy Act of 1974, that was passed in response to concerns about government abuses of power.[4] For nearly 50 years, this legislation has been effective in its primary goal of preventing the U.S. government from centralizing and broadly linking data about individuals across agencies. However, this approach has come at a cost, which is that most government agencies are prevented from freely sharing and linking data across agency boundaries, which in turn hampers agency operational and research efforts.[5] According to one government privacy expert, even when authorized or mandated to share data in limited circumstances, federal agencies are often reluctant to do so due to a myriad of factors, most prominently a lack of adoption of consistent data security standards, as well as difficulties with measuring and assessing privacy risks.[6] To that end, many agencies see promise in adopting technical privacy measures, such as differential privacy, or the creation of synthetic datasets as proxies for actual data, as a necessary precursor for enabling data sharing for both research purposes and interagency goals.[7]

In the nearly 50 years since the Privacy Act's passage, there have been periodic efforts to address the government's approach to data management

Even when authorized or mandated to share data in limited circumstances, federal agencies are often reluctant to do so due to a myriad of factors, most prominently a lack of adoption of consistent data security standards, as well as difficulties with measuring and assessing privacy risks.

while preserving data privacy. Examples include the E-Government Act of 2002, [8] the Confidential Information Protection and Statistical Efficiency Act of 2002,[9] and most recently, the Foundations for Evidence Based Policymaking Act[10] and the National Data Strategy.[11] Most of these efforts have been aimed at sharing government data for statistical analysis and policy evaluation, and the scope of provisions may need to be broadened to support AI research. We view these efforts to be complementary: The NRC should build on these efforts, while bringing increased attention to the compute resources that enable AI development as well as advanced data analysis.

STATUTORY CONSTRAINTS ON DATA SHARING

One vision of the NRC is for it to act as a data warehouse for all government data. But that vision collides with fundamental constraints from laws designed to hamper broad and unconstrained data sharing between U.S. government agencies. Lacking an overarching, comprehensive privacy regime, similar to the European Union's General Data Protection Regulation (GDPR), the US landscape is fragmented between a mix of sector-specific consumer laws and certain government-specific laws, such as the Privacy Act of 1974[12] and limited-scope federal guidance, such as the Fair Information Practice

Principles.[13] In particular, the Privacy Act, which focuses broadly on data collection and usage by federal agencies, and restricts sharing between them, poses challenges to the ambitions of the NRC's goal to make otherwise restricted government datasets more widely available.

Existing efforts, buttressed by such bills as the E-Government Act of 2002 and the Foundations of Evidence-Based Policymaking Act, have attempted to increase access by researchers to government data assets. Yet, these approaches were animated by the primary purposes of policy evaluation, not basic AI research. Nor do they consider any ambitions on the part of agencies themselves to pursue AI research and development.[14]

Application of these laws and regulations to the NRC, in part, hinge on three factors: (1) the institutional form of the NRC, as we discuss in Chapter 4; (2) whether NRC users can invoke the Privacy Act's existing statistical research exception; and (3) whether researchers are accessing data from multiple federal agencies. Here we briefly discuss the legal obligations of federal agencies. Even if the NRC does not take the form of a new standalone federal agency, agencies contributing data will remain subject to these constraints.

THE PRIVACY ACT'S LIMITATIONS AND EXEMPTIONS

The Privacy Act was enacted in response to growing anxiety about digitization, as well as the Watergate scandal during the Nixon presidency. The Act was motivated by concerns about the government's ability to broadly collect data on citizens and centralize it into digital databases, an emergent practice at the time. It is the primary limiting regulation for government data sharing, and has consequences for the NRC and, more directly, for any government agency wishing to share data with the NRC.

Data Linkage

The Privacy Act applies to systems of records, which are defined as "a group of any records under the control of any agency from which information is retrieved by the name of the individual or by some identifying number, symbol, or other identifying particular assigned to the individual."[15] Importantly, the Act places strict limits on "record matching," or linking between agencies, for the purposes of sharing information about individuals.[16] Matching programs are only allowed when there is a written agreement in place between two agencies defining the purpose, legal authority, and the justification for the program; such agreements can last for 18 months, with the option of renewal.[17] These limits were put in place in order to prevent the emergence of a centralized system of records that could track U.S. citizens or permanent residents across multiple government domains, as well as to limit the uses of data for the purposes it was collected. Indeed, while linkage across datasets may be important for AI research,[18] it could potentially enable abuse, surveillance, or the infringement of such rights such as free speech by enabling persecution across the many areas in which a U.S. citizen or resident interacts with the federal system.[19]

Because the restriction on data linkages applies to linkages *between agencies*, the restriction applies in two particular scenarios for the NRC. First, if the NRC is instituted as a federal agency, then agency data-sharing with the NRC would run against the data linkages limitation of the Privacy Act. Second, federal agency staff access to the NRC could raise questions about interagency data linkage under the Privacy Act. However, the recommendation in Chapter 3 is focused on granting agencies streamlined access to the computing resources on the NRC and their own agency data, not to any multi-agency data hosted on the NRC. If the NRC is not designed as a federal agency and does not grant agency members access to interagency data, the Privacy Act's restrictions on data linkages may not apply.

We note that this approach to data management is both unusual and out of step with the private sector, as well as AI research specifically. The ability for both industry[20] and researchers[21] to associate multiple data sources and data points with a specific (anonymized) individual is common practice outside of government. In fact, this limitation is not one that many governments[22] or U.S. states[23] place on their data systems. However, the Privacy Act's restrictions on data linkage remains uncontested, even in the various reform efforts we discuss

below. It is worth noting that the federal government's broad bar against data linkages does incur welfare costs. For example, during the COVID-19 pandemic, the inability to share and link public health data created difficulties tracking the spread and severity of the virus.[24] While projects like Johns Hopkins' Coronavirus Research Center[25] and the COVID Tracking Project[26] attempted to aggregate available data, the lack of data integration slowed important operational and research responses.[27] Other countries, for instance, integrated immigration and travel records to triage cases and prevent hospital outbreaks.[28]

We acknowledge the potential for data linkage to tackle important societal problems without recommending wholesale, unencumbered data linkage. Broad or unrestricted data linkage raises legitimate concerns about both individual privacy and widespread government surveillance,[29] made concrete by the disclosures of government whistle-blower Edward Snowden,[30] among others. An initiative to link Federal Aviation Administration (FAA) data with other agency data for COVID-19 response, for instance, would meet resistance from the Privacy Act. The Task Force should appreciate these tensions and tradeoffs. Indeed, agencies view technical measures for privacy preservation a necessary component of any government data strategy, as methods such as multiparty computation or homomorphic encryption (which we discuss in Chapter 8) may allow for some forms of data linkages between agencies, without violating the Privacy Act.

No Disclosure Without Consent

Another core restriction of the Privacy Act is the "No Disclosure Without Consent" rule, which prohibits disclosure of records to any agency or *person* without prior consent from the individual to whom the record pertains.[31] Because the NRC would disclose federal agency data to researchers (i.e., to "person[s]"), this rule—unlike the restriction on record linkage—is legally relevant and unavoidable.

The Privacy Act, however, contains a number of exceptions to this rule. Most pertinent to the NRC's data-sharing efforts are exemptions for: (1) "routine use"; (2) specified agencies; and (3) statistical research. Under the first exemption, the Privacy Act permits agencies to disclose personally identifiable administrative data when such disclosure is among one of the "routine uses" of the data.[32] A dataset's "routine use" is defined on an agency-to-agency basis, and is simply a specification filed with the Federal Register on the agency's plan to use and share its data.[33] As a result, the more broadly an agency defines "routine use" of its data, the more broadly that agency can share its data with other agencies without disclosure.[34] While courts have limited how broadly an agency can describe "routine uses,"[35] a large number of use cases can still be covered by a short, general statement.[36] Further research should be conducted on the conditions for when data sharing for research purposes constitutes routine use.

Implications for Data Sharing with Researchers

Much will rest on the interpretation of the "statistical research" exception, as applied to AI research. Despite the Privacy Act's constraints on data sharing, researchers have conventionally been able to access data directly from agencies, based on the statistical research exception to the Privacy Act. This exception allows disclosure of records "to a recipient who has provided the agency with advance adequate written assurance that the record will be used solely as a statistical research or reporting record, and the record is to be transferred in a form that is not individually identifiable."[37] Doing so requires either access to an approved research dataset, or for the researcher to negotiate an MOU directly with the agency, a role we suggest the NRC may be able to fill as an intermediary, acting as a negotiating partner to facilitate access requests between multiple researchers and agencies (discussed in Chapter 3).

While the Privacy Act does not define "statistical research," subsequent laws and policies have elaborated on the definition. For example, the E-Government Act defines "statistical purpose" to include the development of technical procedures for the description, estimation, or analysis of the characteristics of groups, without identifying the individuals or organizations that comprise such groups.[38] Meanwhile, a "nonstatistical purpose" includes the use of personally identifiable information for any administrative, regulatory, law enforcement, adjudicative, or other purpose that affects the rights, privileges or benefits of any individual.[39] That is, while researchers may

use personally identifiable data for the broad purpose of analyzing group characteristics, they cannot use such data for targeted purposes to aid agencies with, for instance, specific adjudicative or enforcement functions.

The precise meaning of "statistical purpose," however, remains "obscure and the evaluation criteria may be difficult to locate."[40] Yet, "statistical purpose" may well encompass data sharing for certain AI applications. The Act explicitly designates the Bureau of Labor Statistics, Bureau of Economic Analysis, and the Census Bureau as statistical agencies that have heightened data-sharing powers for statistical purposes.[41] These agencies regularly use AI in conducting their statistical activities.[42] While definitions of AI are themselves contested, statistical research may encapsulate at least some forms of machine learning and AI, if such research analyzes group characteristics[43] and does not identify individuals.

To be sure, the NRC should not enable researchers or agencies to conduct an end run around the Privacy Act. To that end, the NRC will require staff devoted to privacy compliance and oversight to ensure compliance. Key questions regarding individual identifiability, sensitivity of the data, or the potential for linkage and reidentification will need to be assessed by such staff.

Implications for Agency Data Sharing with the NRC

Notwithstanding the above avenues, agencies may nonetheless be reluctant to share data with the NRC and its researchers. Instances abound where federal agencies face constraints to sharing data, even if it is entirely legal or even federally mandated. For example, the Uniform Federal Crime Reporting Act of 1988 requires federal law enforcement agencies to report crime data to the FBI.[44] Yet, no federal agencies appear to have shared their data with the FBI under this law.[45] Similarly, the Census Bureau is enabled by legislation that authorizes it to obtain administrative data from any federal agency and requires it to try to obtain data from other agencies whenever possible.[46] However, the statute does not similarly require the program agencies to provide their data to the Census Bureau. That is, although the Census Bureau is required to ask other agencies for data, those agencies are not required to, and often do not, provide it.[47]

[T]he NRC should not enable researchers or agencies to conduct and end run around the Privacy Act.

Failure to engage in data sharing, even in the face of a statutory authorization, can stem from risk aversion. According to a GAO report, agencies choose not to share data because they tend to be "overly cautious" in their interpretation of federal privacy requirements.[48] Because legal provisions authorizing or mandating data sharing are often ambiguous,[49] agencies may err on the side of caution and choose not to share their data for fear of the downside risk that recipient use of the data may violate privacy or security standards.[50] To make matters worse, because agencies need to devote significant resources to facilitate data sharing, they may simply choose not to prioritize data sharing at all. The lack of resources poses a significant problem; according to a Bipartisan Policy Center study on agency data sharing, about half of agencies cited inadequate funding or inability to hire appropriate staff as their "most critical" barrier to data sharing.[51]

The NRC may overcome these hurdles by clarifying legal provisions, ensuring that the benefits to agencies of data sharing outweigh the risks and costs, and advocating for resources. For instance, O'Hara and Medalia describe how the Census Bureau was able to obtain food stamp and welfare data from state agencies. In the face of ambiguous statutes authorizing the U.S. Department of Agriculture (USDA) and the U.S. Department of Health and Human Services (HHS) to perform data linkages across federally sponsored programs, states originally arrived at different statutory interpretations. Some states agreed to share their data only after (1) the Office of General Counsel at both the USDA and HHS issued a memo clarifying that data sharing with the Census Bureau for statistical purposes was legal and encouraged; and (2) the states were convinced that data sharing would enable evidence building that could help them administer their programs.[52]

Broader data sharing with the NRC that combines multiple agency or external data sources may be facilitated by the passage of additional laws requiring agencies to share their data, subject to specific limitations on how that data is used by the NRC. Even then, the effect of that requirement is hardly a foregone conclusion. More is needed by way of both clarifying the extent to which data sharing is permitted and providing benefits that incentivize agencies to share their data.

Finally, to ensure compliance with the Privacy Act, as well as to facilitate the NRC's role as a data intermediary, the NRC will require a staff of privacy professionals that include positions tasked with legal compliance, oversight, and technical methods expertise. These professionals should build relationships with peers across agencies to facilitate data access.

CASE STUDY: ADMINISTRATIVE DATA RESEARCH UK

Administrative Data Research UK (ADR UK) is a new body, set up in July 2018, to facilitate secure, wide access to linked administrative datasets from across government for the purpose of public research.[53]

ADR UK was set up as a central, coordinating point between four national partnerships—ADR England, ADR Northern Ireland, ADR Scotland, and ADR Wales—as well as the UK-wide national statistics agency, Office for National Statistics (ONS). ADR UK labels itself as a "UK-wide strategic hub": a central point that promotes the use of administrative data for research, engages with government departments to facilitate secure access to data, and funds public good research that uses administrative data.[54]

Funding for ADR UK came from a research council (Economic and Social Research Council, ESRC) and was initially committed from July 2018 to March 2022. A total of £59 million was provided.[55]

ADR UK serves three core functions. First, the promotion of the value and availability of government administrative datasets for research. ADR UK acts as a general advocate for the use of administrative datasets from across the British government. It also acts as a specific driver of research for public good: It has identified specific areas of research that are of pressing policy interest (e.g., "world of work"[56]), and is focusing on creating access to linked datasets for researchers who tackle those priority themes.

The second core function is serving as a coordination point to encourage government data sharing, standards, and linkage of administrative datasets. Especially for its research calls, ADR UK is able to highlight multiple datasets, often spanning different government departments' scope areas that can be linked and used in research. In doing so, ADR UK plays an important role in facilitating research.

KEY TAKEAWAYS

- ***Proactive advocacy for data use and linkage:*** Given the range of agencies and data sources in government, having a single, coordinated voice of advocacy for data use and linkage of public datasets for public good is an important function.
- ***Bringing external talent into government data use:*** ADR UK has two schemes—Research Fellowships and Method Development Grants—that target exceptional, external talent with the intention of building awareness and use of public datasets in cutting-edge research.
- ***Small grant funding to accelerate research methods that use large datasets:*** By putting out calls for research that answers broad themes, ADR UK is able to corral a range of datasets in answering research questions and avoids a single disciplinary focus.

CASE STUDY: ADMINISTRATIVE DATA RESEARCH UK (CONT'D)

Third, ADR UK has a strategic funding approach to further the use of administrative datasets in research that has three categories of funding:

- **Building new research datasets:** ADR UK's Strategic Hub Fund initially solicited invitation-only bids for researchers who would build new research datasets of public significance in the course of their work.[57] These new, research-ready datasets are now accessible to a wide range of researchers.[58]
- **Research Fellowship Schemes:** A major funding focus now is on funding research through competitive open-bid invitations under a Research Fellowship Scheme.[59] Specific researchers are identified through the competition. They are accredited for secure data access and placed right at the heart of government (with 10 Downing Street), with access to linked datasets to answer questions of public significance.[60]
- **Methods Development Grants:** Separately, ADR UK invites research proposals that further methodological progress for the use of large-scale administrative datasets, such that the wider social science community can draw on developed methods in research.[61]

Privacy and Security

The UK's 2017 Digital Economy Act[62] created a legal gateway for research access to secure government data. Deidentified data held by a public authority in connection with the authority's functions could be disclosed for research, under the assurance that individual identities would not be specified.

Any data shared with researchers is anonymized: Personal identifiers are removed, and checks are made to protect against re-identification.[63] A rigorous accreditation process—for both the researcher and proposed research—is undertaken to ensure public benefit. Data access primarily takes place via a secure physical facility, or a secure connection to that facility, provided by ADR UK's constituent partners.[64] There is close monitoring of researcher activity and outputs, and any output is checked before release.[65]

From a researcher's point of view, access to ADR UK datasets requires the following steps:[66]

- Researcher submits proposal for project to ADR UK.
- Project is approved by relevant panels.
- Researcher engages in training and may take assessment (e.g., access to linked data held by ONS required accreditation to ONS' Secure Research Service,[67] and can access data either in person or, where additionally accredited, through remote connection).
- Required data is determined by ADR UK (through one of the four regional partners, or ONS), then ingested by the relevant data center.
- De-identified data is made available through a secure data service (either at the ONS, or one of the four regional partners).
- Researcher conducts analysis; activity and outputs are monitored.
- Outputs are checked for subject privacy. Research serving the public good is published.

COMPLEMENTARY EFFORTS TO IMPROVE THE FEDERAL APPROACH TO DATA MANAGEMENT

The barriers to data sharing created by the Privacy Act have long posed a challenge to researchers interested in using government data to evaluate or inform policy.[68] The policy and statistical research communities, both within and outside the federal government, have engaged in admirable reform efforts to facilitate data sharing for policy evaluation.[69]

The Foundations for Evidence Based Policymaking Act (EBPA) of 2018, which enacted reforms to improve data access for evidence-based decision-making, is a key achievement of these efforts to date. However, several of the provisions in the Act that helped to address some of the barriers to data linking and sharing were not passed by Congress. These provisions—known collectively as the National Secure Data Service (NSDS)—remain a high priority for facilitating further progress for sharing data for research purposes. According to the nonprofit Data Foundation, one of the major supporters of the NSDS, its passage will "create the bridge across the government's decentralized data capabilities with a new entity that jointly maximizes data access responsibilities with confidentiality protections."[70]

The NSDS is envisioned as an independent legal entity within the federal government that would have the legal authority to acquire and use data. However, this authority is currently conceived of as emanating from the EBPA, which focuses on using statistical data for evidence-building purposes. A broader source of authority may be necessary for AI research purposes under the NRC, which may be distinct from agency obligations. One clear area of overlap is the proposal's call for the NSDS to facilitate its own computing resources, which could be harmonized with the compute needs of the NRC. Similar to Chapter 4's discussion of organizational options, NSDS supporters identify a fundamental need for both a reliable funding source as well as thoughtful placement of the NSDS either within an existing agency or as an independent agency or FFRDC. The areas of common ground between the NRC and NSDS, as well as the expertise and momentum behind the proposal, strongly suggest that the NRC engage and coordinate with these efforts.

Another complementary initiative is the Federal Data Strategy (FDS), launched in 2018 by the executive branch and led by the OMB. FDS is a government-wide effort to reform how the entire federal government manages its data. The plan calls out the need for "safe data linkage" through technical privacy techniques,[71] and incorporates a directive from the 2019 Executive Order on Maintaining American Leadership in Artificial Intelligence to "[e]nhance access to high-quality and fully traceable federal data, models, and computing resources to increase the value of such resources for AI R&D, while maintaining safety, security, privacy, and confidentiality protections, consistent with applicable laws and policies."[72] The FDS directs OMB to "identify barriers to access and quality limitations" and to "[p]rovide technical schema formats on inventories," with a focus on open data sources (i.e., non-sensitive or individually identifying data).[73] Datasets identified by this process could be key candidates for populating the NRC.

While both the NSDS and the FDS may promote data sharing, these efforts are presently focused primarily on furthering policy evaluation purposes. Fortunately, there is much overlap and complementarity between these initiatives and the NRC, illustrating the broad importance of more effective mechanisms to share federal data securely and in a privacy-protecting way.

Chapter 6: Technical Privacy and Virtual Data Safe Rooms

KEY TAKEAWAYS

- Technical privacy measures are useful, but not substitutes for securing data privacy through access policies.
- In some instances, the NRC or agencies may wish to make access to data conditional on the use of technical privacy measures.
- Contributing agencies and the NRC should collaborate to determine technical privacy measures based on dataset sensitivity, dataset utility, and equity implications.
- The NRC must have technical privacy staff to administer technical privacy treatments, as well as to support adversarial privacy research.
- The NRC should explore adopting virtual "data safe rooms" that enable researchers to access raw administrative data or microdata in a secure, monitored, and cloud-based environment.

We now discuss the role of technical privacy methods for the NRC. In the past several decades, researchers have devised a variety of computational methods that enable data analysis while preserving privacy. These methods hold considerable promise for enabling the sharing of government data for research purposes. We note at the outset that technical methods are merely one mechanism to strengthen privacy protections. While effective, such methods may be neither sufficient nor universally appropriate. The application of any particular method does not obviate the need to inquire into whether the data itself adheres to articulated privacy standards. The methods discussed here are not "replacements" for the recommendations discussed earlier and never themselves justify the collection of otherwise problematic data.

Use of data from the NRC introduces two threats to individual privacy. The first type involves accidental disclosure by agencies (agency disclosure): An agency uploads a dataset to the NRC which lacks sufficient privacy protection and contains identifying information about an individual. A researcher—either analyzing this dataset alone or in conjunction with other NRC datasets—discovers this information and re-identifies the individual.[1] The second type involves accidental disclosure by researchers (researcher disclosure). Here, a researcher releases products computed on restricted NRC data (e.g., trained machine learning models, publications). However, the released products lack sufficient privacy protection, and an outside consumer of the research product learns sensitive information about an individual or individuals in the original dataset used by the researcher.[2]

We recommend that, due to the infancy and uncertainty surrounding uses of privacy-enhancing technologies, privacy should primarily be approached via access policies to data. While there will be circumstances that suggest, or even mandate, technical treatments, access policies, discussed in Chapter 3, are the primary line of defense: They ensure sensitive datasets are protected by controlling who can access the data. We recommend a tiered access policy, with more sensitive datasets placed in more restricted tiers. For instance, highly restricted access data may correspond to individual health data from the VA, while minimally restricted access data may correspond to ocean measurements from NOAA. Proposals requesting access to highly restricted data would face heightened standards of review, and researchers may be limited to accessing only one restricted access dataset at a time. This approach mirrors current regimes where researchers undergo special training to work with certain types of data.[3]

Technical treatments are a different line of defense: They significantly reduce the chances of deanonymizing a dataset. There are a range of technical methods that can enable analysis while ensuring privacy:

- Techniques like *k*-anonymity and *ℓ*-diversity attempt to offer group-based anonymization by reducing the granularity of individual records in tabular data.[4] While effective in simple settings and easy to implement, both methods are susceptible to attacks by adversaries who possess additional information about the individuals in the dataset.

- One of the most popular techniques is differential privacy,[5] which provides provable guarantees on privacy, even when an adversary possesses additional information about records in the dataset. However, differential privacy requires adding random amounts of statistical "noise" to data and can sometimes compromise the accuracy of data analyses. Although differential privacy has become a point of contention with respect to the Census Bureau's new disclosure avoidance system,[6] the technique remains a powerful defense against bad actors seeking to take advantage of public data for the purposes of re-identification.

- Researchers have also identified other promising methods. Recent work has demonstrated that machine learning can be used to generate "synthetic" datasets, which mirror real world datasets in important ways but consist of entirely synthetic examples.[7] Other work has focused on the incorporation of methods from cryptography, including secure multiparty computation[8] and homomorphic encryption.[9]

Methods that obscure data introduce fundamental tensions with the way machine-learning researchers develop models. For example, when considering questions of algorithmic fairness, in some instances privacy protections can undercut the power to assess whether such a technical method as differential privacy results in demographic disparities, particularly for small subgroups.[10] Similarly, "error analysis"—the study of samples over which a machine-learning model performs poorly—is central to how researchers improve models. It requires understanding the attributes and characteristics of the data in order to better understand the deficiencies of an algorithm. Therefore, such methods as differential privacy, which make raw data more opaque, will invariably impede the process of error analysis. Synthetic data typically captures relationships between variables only if those relationships have been intentionally included in the statistical model that generated the data,[11] and thus, may be poorly suited to certain AI models that discover unanticipated relationships among data. While homomorphic encryption may not require similar assumptions on data structure, existing methods are computationally expensive.

While promising, understanding and applying these methods is an evolving scientific process. The NRC is poised to contribute to their evolution by directly supporting research into their application.

CRITERIA AND PROCESS FOR ADOPTION

The NRC will contain a rich array of datasets, each presenting unique privacy implications over different types of data formats (e.g., individual tabular records, unstructured text, images). Including a dataset on the NRC raises a question of choice: Which technical privacy treatment should be applied (e.g., *k*-anonymity vs. differential privacy), and how should it be applied? This question often requires technical determinations about different algorithmic settings, but such technical choices can also have important substantive consequences.[12]

First, we recommend that these determinations are made with respect to the following factors:

- **Dataset sensitivity:** Different datasets will pose privacy risks that range in type and magnitude. Health records, for instance, are more sensitive than weather patterns. The privacy method chosen should reflect this sensitivity. As we discuss in Chapter 3, these privacy methods should correspond and be tiered

to the appropriate FedRAMP classification for the dataset.

- **Dataset utility:** As discussed above, applying a privacy method can distort the original data, diminishing the accuracy and utility of analysis. Because different methods affect different levels of distortion, the choice of method should be informed by the perceived utility of the data. High-utility datasets—where accurate analyses are highly important (e.g., medical diagnostic tools)—may necessitate methods that produce less distortion.

- **Equity:** Certain privacy measures can disproportionately impact underrepresented subgroups in the data.[13] In determining which method to apply, the presence of sensitive subgroups and their relation to the objectives of the dataset should be evaluated.

For any given dataset, we recommend that agencies providing the data collaborate with NRC staff to identify and recommend any privacy treatments. Originating agencies and NRC staff will possess domain and research expertise to make evaluations on the balance of privacy, utility, and equity, but agencies should consult with NRC staff and researchers on the most appropriate treatments. Given the cost of review, such privacy treatments should be much less widely considered for low-risk datasets.

VIRTUAL DATA SAFE ROOMS

For individual research proposals that would be greatly hampered by technical privacy measures, the NRC should explore the use of virtual "data-safe rooms" that enable researchers to access raw administrative microdata in a secure, monitored environment. Currently, the Census Bureau implements these safe rooms in physical locations and moderates access to raw interagency data through its network of Federal Statistical Research Data Centers (FSRDCs). However, the NRC should not adopt the FSRDC model wholesale. Indeed, the barriers to using FSRDCs are high, and "only the most persistent researchers are successful."[14] For instance, applying for access and gaining approval to use an FSRDC takes at least six months, requires obtaining "Special Sworn Status," which involves a Level Two security clearance, and is limited to applicants who are either U.S. citizens or have been U.S. residents for three years.[15] To further complicate matters, agencies have different review and approval processes for research projects that wish to access agency data using an FSRDC.[16] Finally, even after approval is granted, researchers can only access the data in person by going to secure locations, such as the FSRDC itself.[17]

To be clear, some of these restrictions are unique to the Census Bureau. U.S. law provides that any Census datasets that do not fully protect confidentiality may only be used by Census staff.[18] Researchers trying to access such data therefore must go through the rigorous process of becoming a sworn Census contractor. The extent to which these restrictions apply to the NRC will depend on whether the NRC institutionally houses itself in the Census Bureau, which we ultimately do not recommend.[19] Other problems, however, such as the lack of interagency uniformity in granting access to datasets is not a problem unique to Census, but a common problem throughout the federal government (see Chapter 3).

Another common problem—not necessarily tied to FSRDCs or the Census Bureau—is the use of a physical data room to access raw microdata. The NRC should explore a virtual safe room model, whereby researchers can *remotely* access such microdata. For instance, in the private sector, the nonpartisan and objective research organization, NORC, located at the University of Chicago, is a confidential, protected environment where authorized researchers can securely store, access, and analyze sensitive microdata remotely.[20] Some federal government agencies have also implemented their own virtual data safe rooms. The Center for Medicare and Medicaid Services' Virtual Research Data Center (VRDC), for instance, grants researchers direct access to approved data files through a Virtual Private Network.[21] In a 2019 Request for Information (RFI), the National Institutes of Health also solicited input for its own administrative data enclave and whether such an enclave should be physical or virtual.[22] As articulated in responses to the RFI from the American Society for Biochemistry and Molecular Biology and the Federation for of American

Societies for Experimental Biology, a virtual enclave would greatly facilitate researcher access to data and can be designed and administered in a way to preserve privacy and security.[23]

A National Research *Cloud* cannot function effectively if access to certain datasets is ultimately tied to a National Research *Room*.

CASE STUDY: CALIFORNIA POLICY LAB

The California Policy Lab (CPL) is a University of California research institute that provides research and data support to help California state and local governments craft evidence-based public policy.[24] CPL offers a variety of services to governments, including data analysis services and secure infrastructure for hosting and linking the vast amounts of data collected by government entities.[25] These services help bridge the gap between academia and government by helping policymakers gain access to researchers and providing researchers a secure way to access administrative data. CPL aims to build trusting partnerships with government entities and enable them to make empirically supported policy decisions.

CPL enters data-use agreements with various government entities around California, including, for example, the California Department of Public Health and Los Angeles Homeless Services Authority.[26] These agreements allow CPL to store administrative data in a linkable format, promoting broad longitudinal analyses across various public sector domains.

To help manage the requirements of the various data-use agreements and simplify compliance, CPL applies the strictest requirements for any individual data across all data it stores.[27] Each set of administrative data is thus subject to strict technical restrictions and thorough audits.[28] CPL manages the data in an on-premises data hub at UCLA. This data hub uses "virtual enclaves" modeled after air-gapped clean rooms typically used for sensitive government data.[29] Virtual enclaves are virtual machines that forbid any outbound connections.

CPL creates a new virtual enclave for each research project and only gives specific researchers access to specific datasets for each project.[30] Researchers can only work with the data in the enclave and can only use tools provided in the environment. Data access processes vary, based on the requirements for the government entities, and most of CPL's data-use agreements are purpose limited and thus require approval from the relevant government entity before being used in a project.[31]

KEY TAKEAWAYS

- ***Virtual enclaves:*** The CPL heavily utilizes secure virtual enclaves for researchers to access, work with, and perform data linkages across sensitive datasets.
- ***Acting as an intermediary:*** CPL facilitates and streamlines access to administrative data by acting as an intermediary between researchers and relevant state agencies.

Generally, CPL helps researchers understand how to gain access to various types of administrative data. For some datasets, CPL has formalized applications on its website.[32] CPL prescreens project proposals and sends promising projects to its government partners for final approval. Researchers then conduct these approved projects on CPL's secure infrastructure. For other datasets without formalized access processes, CPL directs researchers toward individuals within the government entities.[33] CPL can then take over management of approved projects that aim to use data stored on its hub under their standing data-use agreements. Alternatively, the government entities and researchers themselves may craft new data-use agreements for specific projects.

IMPLICATIONS FOR THE NRC

DEDICATED STAFF

As discussed above, it will be critical for the NRC to maintain a dedicated professional staff who specialize in privacy technologies. First, not all agencies or departments that seek to place data into the NRC will have the expertise to both determine the privacy method that meets data utility expectations and data privacy demands, and apply it to the dataset of interest. Specialized NRC staff will be essential to assisting such agencies and departments. Second, even where agencies and departments do possess the requisite expertise, NRC staff will bring a unique perspective from their collaborations across the government. Where a specific department's staff may only foresee risks specific to the dataset, NRC staff will be able to foresee instances where the presence of other data in the NRC may raise other concerns. In fact, by working with Affiliated Government Agencies and agency representatives, the NRC staff can also help these agencies internalize such benefits as helping them understand the full range of privacy risks with respect to their data.[34] Such collaborative governance will be necessary to ensure that privacy assessments consider the full implications of access and privacy technologies. Finally, it must not be overlooked that while data management in general requires technical expertise, these various privacy-enhancing technologies also require very specific, highly skilled expertise. Using synthetic data sets as an example, NRC staff could be asked to build synthetic data on an agency's behalf, or need to validate the work performed at an agency to ensure it is done properly and well. Whatever the task, there are cascading effects downstream through the research ecosystem if not carefully managed and executed.

A FOCUS ON EVALUATING AND RESEARCHING PRIVACY-ENHANCING TECHNOLOGIES

It will be necessary to continually evaluate the state of privacy protections on the NRC, either by NRC staff members or by supporting privacy and security researchers at academic institutions. Technical privacy and security research is by nature adversarial: Researchers adopt the posture of adversaries in order to probe the weaknesses of a system/dataset. In the context of the NRC, this will require simulating attacks as researchers try to reidentify individuals within specific NRC datasets. This type of research is necessary to advance the field, and the NRC may be specially positioned to support a research center devoted to researching privacy-enhancing technologies. Doing so would allow the research community to build stronger privacy methods to ensure anonymity, identify flaws, and self-regulate an evolving data ecosystem.

> A National Research *Cloud* cannot function effectively if access to certain datasets is ultimately tied to a National Research *Room*.

Chapter 7: Safeguards for Ethical Research

The pace of advances in AI has sparked ample debate about the principles that should govern its development and implementation. Despite the technology's promise for economic growth and social benefits, AI also poses serious ethical and societal risks. For example, studies have demonstrated AI systems can propagate disinformation,[1] harm labor and employment,[2] demonstrate algorithmic bias along age, gender, race, and disability,[3] and perpetuate systemic inequalities.[4]

This chapter considers how the NRC should ensure its resources are deployed responsibly and ethically. A growing body of research on AI fairness, accountability, and transparency has raised serious and legitimate questions about the values implicated by AI research and its impact on society.[5] The NRC's focus on increasing access to sources of public data and fostering noncommercial AI research is intended to help address these concerns by enabling broader opportunities for academic research. At the same time, broadening access to resources is not enough to assure that academic AI research does not exacerbate existing inequalities or perpetuate systematic biases. In addition, the NRC must also be prepared to handle and act upon complaints of unethical research practices by researchers.

While there is an abundance of proposed ethics frameworks for AI (see Appendix C for those published by federal agencies), there is not a set of accepted principles enshrined into law, like the Common Rule for human subjects research, that clearly establishes the boundaries for ethical research with AI.[6] Lacking such guidance, a core question for the NRC is how to institutionalize the consideration of ethical concerns. This chapter starts by discussing two potential approaches for research proposals: *ex ante* review at the proposal stage for access to NRC resources (e.g., compute, dataset), and *ex post* review after research has concluded. Separately, we discuss guidance for the NRC on issues related to research practices. One of the virtues of starting with access by Principal Investigator (PI) status (Chapter 2) is that researchers will (a) often have undergone baseline training by their home institutions in research compliance, privacy, data security, and practices for research using human subjects; and (b) be subject to research standards and peer review (e.g., through IRB review when applicable). These mechanisms are insufficient to cover many AI research projects, such as when human subjects review is deemed inapplicable. Thus, we tailor our recommendations to the institutional design of the NRC.

First, we recommend that the NRC require including an ethics impact statement for PIs requesting access beyond base-level compute, or for research using restricted datasets. This provides a layer of ethical review for the highest resource projects that are already required to undergo a custom application process. Second, for other categories of research (e.g., research conducted under base-level compute access, where no custom review is

KEY TAKEAWAYS

- Researchers requesting access to compute beyond the default allocation and/or restricted data (i.e., those undergoing a custom application process) should be required to provide an ethics impact statement as part of their application.
- The NRC should establish a process to handle complaints about unethical research practices or outputs.
- Eligibility based on Principal Investigator status will ensure some review under the Common Rule as well as through peer review, but we recommend universities consider more comprehensive models for assessing the ethical implications of AI research.

contemplated), we recommend that the NRC establish a process for handling complaints that may arise out of unethical research practices and outputs. Third, given the limitations of the prior mechanisms, we recommend the exploration of a range of measures to address ethical concerns in AI compute, such as the approach taken by the National Institutes of Health to incentivize the embedding of bioethics in ongoing research.

ETHICS REVIEW MECHANISMS

EX ANTE

Ex ante review assesses research yet to be performed.[7] Funding agencies and research councils worldwide rely on ex ante peer reviews to evaluate the intellectual merit and potential societal impact of research proposals, based on set criteria.[8] Institutional Review Boards (IRBs) commonly assess academic research involving human subjects prior to its initiation.[9] However, much AI-related research may not fall under IRB oversight, as the research may not use human subjects or rely on existing data (not collected by the proposers) about people that is publicly available,[10] used with permission from the party that collected the data, or is anonymized. Potential ethical issues may, therefore, escape IRB review.[11]

Creating an across-the-board ex ante ethics review process, however, would be challenging. First, as we discuss in Chapter Two, we recommend against case-by-case review for all PI requests for access to NRC compute and data, as such a process would require substantial administrative overhead. At the stage when researchers are simply applying for compute access, the research may be so varied and early stage, that there is not much concrete to review. And to the extent that every PI would require project-specific review, such a process would be onerous.

Second, ex ante review is unlikely to grapple with the many ethical implications of design decisions that take place after research commences.[12] Research design can change substantially from initial proposals as projects progress. Ex ante review could identify some concerns, but unlikely all.[13] The nature of machine learning is inherently uncertain—and predictions can be challenging to explain—as well as highly dependent on the data used to build and train models.[14] Ex ante proposal review alone may not be sufficient to identify biased outcomes, and may in fact require extensive documentation and review of the data used in a specific project to assess with any reliability.[15]

Third, there are unique academic speech concerns about government assessment of research. Authorizing the government to conduct an ethics review (separate from IRB review under the Common Rule, which is typically delegated to academic institutions) with vague standards may implicate academic speech concerns, as well as subject proposals to politically driven evaluation that can shift from administration to administration.

If the NRC were to create a process for ex ante review of research proposals for ethical concerns, such a board would likely need to be composed of scientific and ethics experts, similar to how the NSF conducts their process, though perhaps with the addition of members from civil society organizations that focus on countering AI harms. The NSF convenes groups of experts from academia, industry, private companies, and government agencies as peer reviewers, led by NSF program officers and division directors.[16] However, the scope and range of NRC research proposals are likely to be both broad, and highly interdisciplinary in nature, making ethics assessments challenging.

EX POST

Ex post evaluations provide an assessment after research has concluded.[17] In academia, researchers submit research results to journals or conferences for ex post peer review; it is at this pre-publication stage that ethical issues not identified by ex ante processes may be surfaced by reviewers or editors. In the public sector, for example, the Privacy and Civil Liberties Oversight Board (PCLOB) conducts ex post reviews on counterterrorism practices by executive branch departments and agencies to ensure they are consistent with governing laws, regulations, and policies regarding privacy and civil liberties.[18] PCLOB has also recently begun to evaluate the

use of new technologies in foreign intelligence collection and analysis,[19] and to identify legislative proposals that strengthen its oversight of AI for counterterrorism.[20]

RECOMMENDATIONS

While we do not recommend across-the-board ex ante review of research proposals, we do recommend that the NRC establish a process to handle complaints about ethical research practices and outputs. On that point, we recommend the NRC collaborate with the Office of Research Integrity (ORI) at the Department of Health and Human Services to model their processes and procedures for managing issues of research misconduct.[21] The ORI has substantial experience overseeing concerns about ethical research practices. Parties could petition the NRC to revoke access when research is shown to manifestly violate general ethical research standards or practices applicable to a researcher's disciplinary domain. We note that the NRC may want to adopt a high standard for such a violation, given the academic speech considerations. For example, federal agencies or external parties that wish to revoke compute or data access from PIs would need to file a written complaint with supporting evidence. Decisions to revoke access should require input from NRC executive leadership and legal counsel.

For PIs requesting access beyond base-level compute or for restricted datasets, we recommend requiring the completion of ethics impact statements to be submitted with research proposals. A recent proposal to address the lack of "widely applied professional ethical and societal review processes" in computing piloted such a requirement in a grant process, requiring a description of the potential social and ethical impacts and mitigation efforts by researchers.[22] We limit this approach to proposals for compute access beyond default allocation or requests for access to restricted datasets, as the administrability concerns are weaker for researchers who are already applying for compute or data access beyond the default levels. For those applications, a review process of a specific proposal will already occur by an external review panel of experts (Chapter 2), and, much like the NSF requires statements of "Broader Impacts;"[23] statements about the ethical considerations of the work could easily be included. It is important to note that ethics impact statements would be only one component of NRC applications and should be weighed in conjunction with other application materials. In addition to requiring researchers to carefully think through and document the potential impacts of their own work, the statements may also serve as useful documentation of potential negative impacts and be of use to NRC staff when determining whether to provide access to specific types of data. Such assessments may also be helpful for journals, conferences, or universities addressing ex post concerns about ethical impacts.

> [E]mbedded ethics approaches may . . . identify[] and address[] issues as the research proceeds, in contrast to ex ante review, where it may be too early to spot an issue, and ex post review, which may be too late.

Next, we recommend that the NRC employ a professional staff devoted to ethics oversight, similar to what we propose regarding data privacy in Chapters 5 and 6. In addition to staff devoted to handling legal compliance issues, the NRC needs staff with specialized training in AI ethics (as well as expertise in other subdomains) to provide expert internal consulting to NRC applicants, as well as to aid in evaluating ethics impact statements. Similarly, data privacy experts can identify ethical privacy issues specifically related to data, such as whether consent has been properly obtained and documented. To ensure that decisions are based on the merits, the NRC staff overseeing these issues must operate independently of other federal agencies and be insulated from political interference.

We acknowledge that these ethics review mechanisms may not identify all instances where researchers use the NRC in a way to conduct research that raises ethical questions. Few review mechanisms could, particularly in light of the considerable ambiguity present in ethics standards (see Appendix C). Nonetheless, these mechanisms can augment key academic checkpoints (IRB review and peer review) in an administrable fashion that does not raise serious concerns about academic speech.

Lastly, we recommend that non-NRC parties explore a range of measures to address ethical concerns in AI compute. These may include an ethics review process or approaches widely deployed in bioethics by the National Institutes of Health, namely to incentivize the embedding of ethicists in research projects.[24] Such embedded ethics approaches may have the particular advantage of identifying and addressing issues as the research proceeds, in contrast to ex ante review, where it may be too early to spot an issue, and ex post review, which may be too late. We expect this to be an active area of inquiry as new approaches are validated. The NRC, potentially in conjunction with the NSF, should consider offering funding for projects that embed ethics domain experts into teams, in order to support this proposal.

Chapter 8: Managing Cybersecurity Risks

While the NRC has the potential to level the playing field for AI research, it will also create an alluring target for a vast array of bad actors.

While the NRC has the potential to level the playing field for AI research, it will also create an alluring target for a vast array of bad actors.

Cybersecurity—the effort to protect systems against incidents that may compromise operations or cause harm to relevant assets and parties—will be a critical focus of the NRC. It will require a cybersecurity framework that manages potential incidents throughout their *lifecycle,* spanning: (1) preparation; (2) detection and analysis; (3) containment, eradication, and recovery; and (4) post-incident activity, which collectively encompasses incident monitoring, detection, recovery, and reporting.[1] Effective cybersecurity practices complement risk assessment based on impact, immediacy, and likelihood, and will help gain the trust of users and thwart subversion and interference from foreign actors or other adversarial parties. Careful administrative design of the NRC with cybersecurity at the forefront will set a high standard as information systems become more central to our national infrastructure.

In this chapter we address these cybersecurity concerns. We first provide an overview of common types of vulnerabilities and attacks, and assess their relevance to the NRC. Next, we provide an overview of the federal government's regulatory landscape, as it pertains to cybersecurity, with a special focus on the FISMA and FedRAMP frameworks. Finally, we close with a discussion of the security and system design measures best suited to ensure that the integrity of the NRC is not compromised.

KEY TAKEAWAYS

- Deterring malicious actors from attacking the NRC will require more than adhering to current FISMA and FedRAMP standards.
- The NRC should centralize security responsibilities for datasets with the program's staff rather than deferring to originating agencies.
- Technical measures the NRC should investigate include confidential clouds, federated learning, and cryptography-based measures such as homomorphic encryption and secure multiparty computation.

MOTIVATIONS FOR POTENTIAL ATTACKS

Possible attacks against the NRC could take a number of approaches, each of which would entail substantial consequences for the NRC.[2] First, adversaries could launch an attack against the NRC with the intention of disrupting its operations or its ability to aid research. For example, adversaries could attack the NRC's infrastructure directly by disabling or interfering with NRC servers. As a result, researchers would be unable to access NRC servers or effectively utilize them. By launching such attacks, adversaries may throttle the NRC, thereby raising costs for the federal government.[3] Alternatively, adversaries could seek to attack specific research projects on the NRC,

thereby slowing the pace of that research or compromising the quality of the research findings. They may also initiate "data-poisoning" attacks on NRC datasets, thereby compromising the quality of research findings.

Second, bad actors could also launch cyber operations against the NRC, intending to steal computational resources. In this case, the purpose would not be to disrupt the NRC, but to repurpose computational power toward illicit purposes (e.g., cryptocurrency mining).[4] For instance, individuals could pretend to be researchers, claiming to use cloud credits for legitimate research purposes while actually using them for alternative ends. Individuals could also infiltrate the NRC's network, siphoning off computational resources from other projects and reducing the functionality for legitimate users.

Third, adversaries might pose a threat to the NRC out of a desire to steal or make use of the data and research products housed within the system. The NRC promises to be an attractive target because it will house data from a range of different agencies. If an adversary wanted to steal equivalent data from the agencies themselves, they would need to break into each agency independently. However, the potential combination of datasets on the NRC, including researcher-owned datasets, may increase the potential gains from accessing this information. Additionally, adversaries may attempt to break into the NRC in order to steal products generated by NRC researchers. This could include trained machine-learning models or specific research findings.

Relatedly, bad actors could determine that executing intrusions into the NRC is an effective way to target Affiliated Government Agencies. Because a participation incentive for agencies is the computing support that the NRC will offer, one of the biggest cyber risks is of malicious actors attempting to use the NRC to hack into their systems. For that reason, the cybersecurity risk to the government may be substantial. On the other hand, as we discussed in Chapter 3, the NRC also presents an opportunity to enhance and harmonize security standards compliance, as agencies move into the cloud.

A range of other motivations may exist. Successful operations against the NRC, as a federal entity, would carry symbolic value and capture attention. Ransomware attacks could result in significant payoffs. The NRC could also be a target for espionage, both on the part of nation-state actors seeking to acquire sensitive datasets (e.g., energy grid infrastructure) and on the part of private sector entities looking to steal intellectual property or to monitor the latest technological advances.

If successful, any attack could undermine the NRC. For example, researchers would be deterred from using the NRC and may invest their efforts in alternate private clouds. This could occur because researchers believe the NRC would be ineffective to use (e.g., on account of frequent server outages), or because they believe their research products would be inadequately protected. Federal agencies and departments could be deterred from entrusting the NRC with sensitive datasets. Federal entities could risk embarrassment and face obstacles executing their policy objectives if datasets were accidentally leaked. If the NRC is insufficiently secure, such entities may choose to avoid sharing data altogether.

FISMA, FEDRAMP, AND EXISTING FEDERAL STANDARDS

As a federal entity, the NRC will be subject to federal standards and regulations. In this section, we provide a high-level overview of the two most relevant regulations: the Federal Information Systems Management Act (FISMA) and the Federal Risk and Authorization Management Program (FedRAMP).[5] FISMA traditionally applies to non-cloud systems that support a single agency, whereas FedRAMP authorization is required for cloud systems.[6] We finish by discussing critiques of these regulations.

FISMA

The Federal Information Systems Management Act (FISMA) was first passed in 2002, with the purpose of providing a comprehensive framework for ensuring the effectiveness of security controls for federal information systems.[7] The law was later amended in 2014, and has

since been augmented through other individual legislative and executive actions, and our discussion focuses on the collective impact of FISMA compliance regulations.[8]

FISMA applies to all federal agencies, contractors, or other sources that provide information security for information systems that support the operations and assets of the agency.[9] It invests responsibility in several different entities. First, the National Institute of Standards and Technology (NIST) is tasked with developing uniform standards and guidelines for implementing security controls, evaluating the riskiness of different information systems and other methodologies.[10] Second, the Office of Management and Budget (OMB) is tasked with overseeing agency compliance with FISMA and reporting to Congress on the state of FISMA compliance.[11] Third, the Department of Homeland Security is tasked with administering the implementation of agency information security policies and practices.[12] Finally, federal agencies are required to develop and implement a risk-based information security program in compliance with NIST standards and OMB policies.[13] Agencies are further required to conduct periodic assessments to ensure continued efficiency and cost effectiveness.[14]

Several NIST requirements are worth mentioning here. Pursuant to NIST SP 800-18, agencies are required to identify relevant information systems falling under the purview of FISMA. Agencies must also categorize each of these systems into a risk level, following the guidance laid out in FIPS 199 and NIST 800-60.[15] NIST 800-53 outlines both the security controls that agencies should follow and the manner in which agencies should conduct risk assessments.[16] Agencies must further summarize both the security requirements and implemented controls in "security plans," as outlined in NIST 800-18.[17] Finally, organization officials are required to conduct annual security reviews in accordance with NIST 800-37.

FEDRAMP

In the late 2000s, federal agencies began expressing security concerns as a barrier to cloud computing adoption.[18] In response, Congress passed the 2011 Federal Risk and Authorization Management Program (FedRAMP) to provide a cost-effective, risk-based approach for the adoption and use of cloud services by the federal government.[19] FedRAMP approval is exempted where: (i) the cloud is private to the agency; (ii) the cloud is physically located within a federal facility; and (iii) the agency is not providing cloud services from the cloud-based information system to any external entities.[20] Like FISMA, FedRAMP security requirements are governed by NIST standards, including NIST SP 800-53, FIPS 199, NIST 800-37, and others.[21] However, unlike FISMA, FedRAMP's two tracks to receiving an authority-to-operate means that vendors working with multiple agencies do not necessarily need to undergo the full approval process with each agency. This means that cloud services providers and agencies alike are able to save significant time and money.

CRITICISMS OF FISMA AND FEDRAMP

These regulations are not without fault. Most notably, critics point to the fact that despite their existence, cyber intrusions on government infrastructure are common and accelerating.[22] A 2019 report by the U.S. Senate Committee on Homeland Security and Governmental Affairs investigating eight agencies noted that the federal government is failing its legislative mandate from FISMA.[23] The errors identified included a failure to protect personally identifiable information, inadequate IT documentation, poor remediation of bugs, a failure to upgrade legacy systems, and inadequate authority vested in agency chief information officers.[24] Reports by the Government Accountability Office (GAO) have reached similar conclusions.[25] In turn, some have criticized the government's approach to cybersecurity wholesale, arguing it places too much emphasis on merely detecting intrusions.[26] They argue for a framework of "zero trust," which assumes that intruders will penetrate a network and instead focus on security controls limiting the ability of those intruders to navigate the network.[27]

FedRAMP faces its own criticisms. A recent study noted that securing authorization can be time-consuming and expensive—taking up to two years and costing millions of dollars in some cases.[28] Even though parts of FedRAMP are designed to be reusable across agencies, agencies often delay the process by imposing separate,

additional requirements. A variety of reasons for these deficiencies have been noted, including an understaffed Joint Authorization Board, a lack of trust between agencies with regards to Authorization to Operate (ATOs), and an overly complex authorization process that leads to errors by agencies and Cloud Services Providers.[29] Proposed recommendations to address these deficiencies include increased funding for FedRAMP's Joint Authorization Board, incentives to encourage reuse of ATOs, and mechanisms to improve the efficiency of the authorization process.[30]

On May 12, 2021, the Biden administration released an Executive Order (EO) on Improving the Nation's Cybersecurity,[31] and OMB published a draft federal strategy for public comment on September 7, 2021.[32] Signed in the aftermath of the breach of the software vendor SolarWinds, and the ransomware attack on Colonial Pipeline, the EO presents several new initiatives. First, it calls on the federal government to embrace "zero-trust architecture" and improve post-attack investigation processes. Second, it seeks to improve collaboration between the public and private sectors by improving disclosure requirements and establishing a private-public Cybersecurity Safety Review Board (modeled after the National Transportation Safety Board). Finally, it seeks a more cohesive government-wide approach to cybersecurity, calling for the creation of a playbook to standardize cyber response across federal agencies, alongside a government-wide detection and response system for attacks.

Though it is too soon to determine whether the EO and the proposed strategy will be effective, it appears to address deficiencies identified in the existing landscape. It seeks to improve documentation and responsiveness to attacks and suggests a shift in cybersecurity thinking. It is unclear, however, whether it will address the underlying procurement issues and lack of interagency trust that critics believe have hampered the effectiveness of FedRAMP. But given the potential for highly sensitive data to be stored on the NRC, embracing a zero-trust architecture at the outset is a crucial consideration for ensuring its integrity.

NRC SECURITY STANDARDS AND SYSTEM DESIGN MEASURES

Here, we present recommendations on cybersecurity policy for the NRC informed by the landscape of the existing federal regulations and unique considerations that a national research cloud will pose.

PROCESS FOR RISK AND SECURITY DETERMINATIONS

Under the current regulatory landscape, agencies are responsible for determining the appropriate risk categorizations and security controls for the datasets located on their servers. However, this raises a potential challenge as agencies begin to share their data with the NRC—making it unclear who will maintain authority for categorizing the risk of these datasets and determining appropriate security controls.

On the one hand, agencies themselves could continue to retain discretion over the security classification and controls for datasets they place into the NRC. In this decentralized approach, much of the security responsibilities assigned by FISMA would remain with the agencies, irrespective of whether the data existed on NRC servers. On the other hand, the NRC could take responsibility for all security decisions. Datasets added to the NRC would then be classified according to the NRC's assessment of risk, and protected with controls that the NRC staff deems appropriate. This approach "centralizes" security responsibilities by vesting it with the NRC after the onetime negotiation for each dataset.

Though both approaches have their merits, we recommend the centralized approach for several reasons. First, the centralized approach ensures internal uniformity. The paradox of federal cybersecurity regulation is that although NIST has articulated a set of standards pertaining to risk and controls, agencies interpret these standards differently, leading to discrepancies in implementation and classification across the federal government. Following each agency's security classifications for data on the NRC would produce unnecessarily complex and incoherent classifications for a single system. This threatens to diminish the usability of the NRC, and the added

complexity could arguably weaken security by increasing the likelihood of errors. Permitting the NRC to impose its own classifications allows for uniformity within the NRC and alignment with the access tiers suggested in Chapter 3 of this White Paper. This approach may also simplify managing security practices across a potential mix of cloud compute providers.

Second, the NRC represents a valuable opportunity to harmonize federal cybersecurity standards across different agencies. The assessments and implementations adopted by the NRC must generalize to the full diversity of federal datasets. Hence, the NRC's practices can serve as a template for NIST's guidelines, which any agency is free to adopt.

Third, the centralized approach will remove hurdles for data sharing. Security concerns often impede agency data sharing. In a scheme where agencies retain control over all security determinations, agencies could demand security classifications that are excessively high or impractical to implement. The centralized approach would place the burden on agencies to articulate with specificity why the NRC's security policies or classification guidelines are inadequate for a particular dataset.

Finally, researchers should also have a voice in determining the appropriate security controls, since a public resource of this magnitude that cannot attract users is bound to fail. As security controls implicate usability, the NRC should not opt for controls that substantially inhibit or disincentivize researchers from leveraging its resources. The NRC needs to strike the right balance between usability and security.

TECHNICAL CONSIDERATIONS

The federal government already possesses a range of technical options and countermeasures to cyberattacks. Cybersecurity threats and defenses are, of course, actively evolving, so we discuss these only as a starting point—robust, long-term cybersecurity comes through continued vigilance and prioritization that recognizes the shifting nature of the field.

DATA STORAGE

Data storage mechanisms should ensure proper protection from outside access. Encryption can be used to protect sensitive data at rest, to be later unencrypted when needed. Physical isolation through air-gapped environments is another design feature that can remove the possibility of wireless network interfaces from being used to connect the data to malicious outside threats. However, even air gapping is not a foolproof solution: There are ways to "jump" air gaps such as through hiding in USB thumb drives (which is allegedly how the Stuxnet malware famously compromised Iranian nuclear centrifuges).[33] More recent attacks bypass the need for electronic transmission altogether by leveraging other signals that leak data, such as FM frequencies, audio, heat, light, and magnetic fields. These kinds of threats bring home the need for a comprehensive and evolving approach to cybersecurity.

NETWORKING PROTOCOLS

Data packets sent over networks are transmitted according to a set of internationally standardized internet protocols. Following the Open Systems Interconnection (OSI) model, the conceptual layers involved in computer networking can be categorized into seven dimensions: physical, data link, network, transport, session, presentation, and application layers.[34]

RUNTIME SECURITY

When considering runtime security technologies, three design features that are relevant for the cloud environments are the use of confidential clouds, federated learning, and cryptography-based measures such as homomorphic encryption and secure multiparty computation. A growing number of vendors offer "confidential cloud" options as an emerging technical solution to fully cyber secure cloud computation that is secure throughout execution.[35] Confidential clouds offer high-security, end-to-end, isolated operation by executing workloads within trusted execution environments. For example, virtualization enables an operating system to run another operating system within it as a virtual environment with additional firewall or other network

barriers, effectively simulating another device within the host computer.

DISTRIBUTED COMPUTING AND FEDERATED LEARNING

Another computing paradigm, known as distributed computing or federated learning, considers situations where multiple parties have individual shards of data they are interested in leveraging in aggregate, without sharing outright. Federated learning addresses this situation, for example, demonstrating how users' mobile phones can send information—possibly differentially private—to central servers without exposing the precise details of any one individual's information. A second scenario more relevant to the large-scale decentralized nature of the NRC is distributed computing—in which many institutions collectively share compute, akin in some respects to crowd-sourced computing. These approaches enable multiple parties to leverage existing computational infrastructure, while retaining some guarantees on privacy.

CRYPTOGRAPHY-BASED MEASURES

Finally, there are two types of cryptography-based measures worth noting.

Cryptography researchers have developed ways of computing mathematical operations over *encrypted* data, known as homomorphic encryption. This impressive feat has valuable implications because it obviates the need for decryption, which can potentially expose the intermediate values of computation, and grant access to public and secret encryption keys during computation. Initially, only partially homomorphic encryption schemes that supported limited arithmetic operations like addition and multiplication were possible. But fully homomorphic encryption schemes have recently been developed that enable what is known as "arbitrary" computation for promising use cases in predictive medicine and machine learning. That said, standardization is still underway to broader adoption, and homomorphic encryption (by design) is malleable—a property in cryptography that is usually undesirable as it allows attackers to modify encrypted ciphertexts without needing to know their decrypted value. These and other limitations of any technical approach are worth taking into account when considering which technologies to adopt and for what purpose.

Complementing the distributed, decentralized computing model discussed throughout this White Paper is the subfield known as secure multiparty computation (also known as privacy-preserving computation), which presents methods for multiple parties to jointly compute a function over all their respective inputs, while keeping those inputs private from other parties. These methods have matured in their origins from a theoretical curiosity to techniques with practical application in studies on tax and education records, cryptographic key management for the cloud, and more.[36] This makes secure multiparty computation methods a potential candidate for applications pertaining to secure, distributed computation.

Ultimately, it will be central for the NRC to continuously learn about the most effective security standards (including such other creative strategies as red teaming or bug bounties[37] to identify vulnerabilities) in this rapidly evolving space.

Chapter 9: Intellectual Property

Who should own the IP rights to outputs developed using NRC resources?[1] When private research is funded, subsidized, or influenced by the federal government, the laws and rules have evolved, so that both the researcher and the government have certain rights in the intellectual property developed under the research. While IP protection is theoretically designed to incentivize research and innovation, some signs indicate that AI researchers in particular are already amenable to sharing the fruits of their research. Indeed, over 2,000 researchers signed a 2018 petition to boycott a new machine intelligence journal started by *Nature*, because it promised to place its articles behind a paywall.[2] The Open Science and Open Research movements have also encouraged AI researchers to make their machine-learning software and algorithms publicly available.[3] Furthermore, as we discuss below, the advancement of techniques like transfer learning depend on researchers being able to distribute the fruits of their research freely.

This chapter surveys the existing IP-sharing agreements between researchers and the government, and explores whether and to what extent the government should retain IP rights over researchers'' outputs, as a condition of using the NRC.[4] While the evidence on optimal IP rights varies, we recommend that: (1) Academic researchers and universities should retain the same IP rights as the Bayh-Dole Act provides for patents developed under federally funded research; (2) The government should retain its copyrights and data rights under the Uniform Guidance, but contract around those rights where applicable to incentivize NRC usage and AI innovation; and (3) The government should consider conditions for requiring researchers to share their research outputs under an open-access license.

[A]cademic researchers and universities should retain the same IP rights as the Bayh-Dole Act provides for patents developed under federally funded research.

PATENTS RIGHTS

A core question is whether NRC users should retain patent rights in inventions supported by the NRC. The Bayh-Dole Act regulates patent rights for inventions developed under federal funding agreements and its applicability depends on the

KEY TAKEAWAYS

- To harmonize with the federal grant process, the NRC should adopt the same approach to allocating patent rights, copyrights, and data rights to NRC users as applies to federal funding agreements.
- The NRC should contract around government intellectual property rights where applicable to incentivize NRC usage and AI innovation.
- The NRC should consider conditions for requiring researchers to share outputs under an open-source license.

nature of NRC access; for instance, if cloud credits are apportioned using federal grants, as described in Chapter 2, they may be considered federal funding agreements.[5] In such cases, Bayh-Dole Act permits researchers to hold the title to the patent and to license the patent rights.[6] However, these patent rights come with certain restrictions: For example, the funding agency has a free, nonexclusive license to use the invention "for or on behalf of the United States," and the agency may use "[m]arch-in rights" to grant additional licenses.[7]

The broader policy question about the government's exercise of its patent rights is whether and how patents stimulate innovation in AI. Some commentators have argued that the U.S. suffers from over-patenting in software,[8] and AI is no exception.[9] The total number of AI patent applications received annually by the U.S. Patent and Trademark Office more than doubled from 30,000 in 2002 to over 60,000 in 2018,[10] and some argue that this proliferation of broad AI patents, especially those filed by commercial companies, is hindering future innovation.[11] In the Bayh-Dole context, researchers have also found that the benefits of university patenting may justify the costs only where industry licensees need exclusivity to justify undertaking the costs of commercialization, as, for instance, in the pharmaceutical context.[12] For the substantial portion of university patenting, including AI, this rationale may not carry much weight.[13]

Some research shows that patents actually may not actually have any net effect on the amount or quality of AI research conducted in the university context. In an empirical study of faculty at the top computer science and electrical engineering universities in the United States, research has found that the prospect of obtaining patent rights to the fruits of their research does not motivate researchers to conduct more or higher-quality research.[14] Eighty-five percent of professors reported that patent rights were not among the top four factors motivating their research activities, and 57 percent of professors reported that they did not know whether or how their university shares licensing revenue with inventors.[15] The patent scheme adopted by the NRC, therefore, may not have a strong influence on researcher adoption.

> The government should . . . consider conditions for requiring NRC researchers to disclose or share their research outputs under an open-access license.

That said, as a practical matter, there is a virtue to treating innovations stemming from NRC usage in a fashion that is consistent with Bayh-Dole. Particularly if cloud credits are awarded through the expansion of programs like NSF CloudBank, it would be confusing to have distinct patent rights out of the research and cloud grant. In addition, many university tech transfer offices appear to have strong preferences for patent rights.[16] To the extent that universities view retaining patent rights as a condition for using the NRC, aligning NRC patent rights with Bayh-Dole may be preferred, but the evidence underpinning this recommendation is not strong.

COPYRIGHT, DATA RIGHTS, AND THE UNIFORM GUIDANCE

The Uniform Guidance (2 C.F.R. § 200) streamlines and consolidates government requirements for receiving and using federal awards to reduce administrative burden.[17] Grants.gov describes it as a "government-wide framework for grants management," a groundwork of rules for federal agencies in administering federal funding.[18] The Uniform Guidance includes provisions on, for instance, cost principles, audit requirements, and requirements for the contents of federal awards.[19]

The Uniform Guidance is applicable to "federal awards,"[20] but IP provisions do not *require* the government to assert their rights over researcher outputs.[21] Whether and how the government allocates its IP rights under the Uniform Guidance is therefore an important question.

This section first covers government copyright and data rights to IP under the Uniform Guidance and

discusses how sharing copyright and data rights might impact the AI innovation landscape. We then examine the extent to which the government should retain its rights to research generated using the NRC. While the evidence is mixed, we ultimately recommend that the government retain its copyrights and data rights under the Uniform Guidance, but contract around those rights where applicable, to incentivize NRC usage and further AI innovation.

COPYRIGHT

Under U.S. copyright law, NRC researchers can obtain copyrights over various aspects of their work. For instance, NRC researchers may wish to copyright the software they used to build the model, since software is considered a literary work under the Copyright Act.[22] Researchers may even obtain copyrights over various aspects of the model, including the choices of training parameters, model architectures, and training labels, if they can show that those choices required creativity.[23] Many scholars have even opined, without reaching consensus, on whether outputs such as text and art that are artificially generated can be copyrighted.[24]

Under the Uniform Guidance,[25] the recipient of federal funds may copyright any work that was developed or acquired under a federal award. However, even if researchers are permitted to maintain copyrights, the federal awarding agency reserves a "royalty-free, nonexclusive and irrevocable right to reproduce, publish, or otherwise use the work for federal purposes, and to authorize others to do so."[26] Notably, this right is limited to "federal purposes," meaning that third parties who acquire licenses to the researchers' copyrighted works cannot use them for exclusively commercial purposes.[27]

It is unclear to what extent copyrights over NRC outputs should be fully vested in the researcher to stimulate basic AI research. One class of AI research and development output that has received significant academic attention has been whether AI-generated creative works, like music from OpenAI's Jukebox,[28] can or should receive copyright protection.[29] However, the technology and copyright community has hardly reached a consensus on whether the public interest in AI research requires granting copyright in these scenarios. On one hand, in a survey of AI scientists, tech policy experts, and copyright scholars, roughly 54 percent of respondents agreed that copyright protection is an important incentive for authors to make their work commercially available, and 63 percent agreed that an increase in the number of commercially available AI-produced works would stimulate further AI growth and research.[30] On the other hand, in the same survey approximately 56 percent of respondents agreed that the U.S. Copyright Office should *deny* copyright protection to creative works produced independently by AI without creative intervention from a human author.[31]

Notwithstanding the prominent debate about copyright over creative works generated by AI models, such works are only a subset of possible copyright protection in the AI context. As discussed above, researchers could theoretically seek additional copyright protection over, among other things, their code, architecture, or model. Here, AI innovation may depend on sharing these copyrightable elements. For instance, transfer learning uses existing ML models and "fine-tunes" those models for a related target task,[32] and various fine-tuning approaches have emerged to perform transfer learning on different classes of tasks.[33]

DATA RIGHTS

Under the Uniform Guidance, when "data" is "produced" under a federal award, the government reserves the right to: (1) obtain, reproduce, publish or otherwise use such data; and (2) authorize others to receive, reproduce, publish or otherwise use such data.[34]

Notably, this does *not* limit the use of such data for federal government purposes. In other words, such data can be promulgated for *any* use. The outstanding question, therefore, is whether this "data," which is not explicitly defined in the Uniform Guidance, covers data generated for AI and machine-learning purposes. Below, we examine two classes of data generated for AI purposes—synthetic data and data labels—and how sharing this data could impact AI innovation.

One class of data generated for AI purposes is synthetic data. Researchers have turned to deep generative models such as Variational Autoencoders[35] and Generative Adversarial Networks[36] to generate synthetic data to train their machine learning models. As noted by the World Intellectual Property Organization, synthetic data is an entirely new class of data that does not fit neatly under existing IP law.[37] While a researcher may seek copyright protection over the subset of synthetic data that is "creative," therefore implicating the copyright provisions of the Uniform Guidance (described above), the broad class of synthetic data, whether "creative" or not, may also implicate the data rights provision. On the one hand, training data is often carefully guarded,[38] so requirements to share synthetic data, which is often used to train AI models, may be a non-starter for NRC users. On the other hand, many scholars have written about the promise of synthetic data to actually *enable* further data sharing by preserving privacy and researchers' trade secrets.[39] In fact, sharing synthetic datasets would spur additional research and innovation in fields such as healthcare, where data sharing has been limited.[40]

Another class of data generated for AI is labeled data, namely data that has been tagged and classified to provide ground truth for supervised machine learning models.[41] While techniques have been developed to decrease the costs associated with data labeling,[42] it nevertheless remains a resource and time-intensive task. For example, Cognilytics Research reports that 25 percent of the total time spent building machine learning models is devoted to data labeling.[43] Researchers using the NRC may therefore seek to protect their investment in data labeling by opting not to share their labels with others, especially if the underlying data is proprietary.[44] However, recognizing the difficulty of data labeling, some researchers have built online platforms for sharing data labels.[45] In the case of ImageTagger, a data labeling and sharing platform for RoboCup Soccer, the developers wanted to solve the problem that no single team, acting alone, could easily build its own high-quality training sets.[46] Similarly, in the NRC's case, the sharing of labeled government data—where labeling may have been augmented by NRC resources[47]—could act as a rising tide that lifts all boats, improving the quality of not only the government data as a training dataset, but also all subsequent research using that data. Furthermore, sharing data labels could be instrumental in conducting bias and fairness of NRC research outputs where necessary, as discussed in Chapter 7.[48]

RETAINING IP RIGHTS IN THE UNIFORM GUIDANCE

As the preceding discussion suggests, sharing AI research output covered by copyrights and data rights could be beneficial to AI innovation. We therefore recommend that the NRC at least retain the same rights to copyrights and data rights as under the Uniform Guidance, yielding several additional benefits. First, similar to our recommendation in Chapter 3 that federal agencies should be allowed to use the NRC's compute resources, retaining the same Uniform Guidance IP allocation scheme could produce welfare benefits by improving government decision-making using AI. For instance, federal agencies can reduce the cost of core governance functions and increase agency efficiency and effectiveness by using data labels shared by NRC researchers or by fine-tuning models generated by NRC researchers. Second, retaining the Uniform Guidance IP allocation scheme would result in more consistency across the federal award landscape. Indeed, as mentioned above in the patent context, it could be confusing to diverge from the Uniform Guidance, especially if the cloud credit grant is apportioned through programs like CloudBank but the research grant is administered as a federal award.

In sum, we recommend that the government at least retain its copyrights and data rights under the Uniform Guidance. However, we also reiterate that the Uniform Guidance serves merely as a helpful framework, not as an immutable rule. Where the Uniform Guidance IP allocation would dissuade researchers from using the NRC or hinder AI innovation in specific scenarios, the government can and should explicitly modify its rights and contract separately with researchers on what rights the government retains, if any.

CONSIDERATIONS FOR OPEN-SOURCING

Should the government go *beyond* its rights and mandate that researchers share their NRC research outputs with others under an open-source license? As an initial matter, we note that agencies can modify the IP allocation schemes under the Uniform Guidance, but not under the Bayh-Dole Act. Some federal agencies supplement and/or replace the IP rights set out in the Uniform Guidance with restrictions that are more specific to the IP being developed for that particular agency or under a specific award.[49] For instance, the Department of Labor requires that intellectual property developed under a federal award must not only comply with the terms specified in the Uniform Guidance, but also be available for open licensing to the public.[50] NSF grantees are also expected to share their data with others.[51] However, the government cannot change the allocation of patent ownership under the Bayh-Dole Act, unless the Act itself is modified or unless the NRC isn't administered as a federal award, rendering the Act inapplicable.

Requiring researchers to open-source their research outputs may be possible, but the considerations around it are complex. On the one hand, an open-source requirement could negatively affect downstream commercialization, given the wide range of potential AI research.[52] While the NRC might protect commercialization to some degree by adopting a restrictive open-source license,[53] the mere divergence from the Uniform Guidance or the Bayh-Dole Act could be confusing for researchers in navigating federal awards and understanding open-source licensing interactions across multiple situations.[54] Furthermore, requiring researchers to share research outputs comes with its own host of privacy and cybersecurity issues.[55] If researchers are permitted to use the NRC to conduct classified research,[56] for instance, then keeping research outputs proprietary would serve the national interest.[57] In this case, however, the NRC should consider limiting any open-source requirement to research that has fewer privacy and security implications.

On the other hand, as discussed, sharing research outputs with other NRC researchers could be beneficial, and many scholars argue that AI researchers should open-source their software to stimulate innovation.[58] A requirement to open-source software code, which can be the subject of both copyrights and patent rights,[59] may contravene Bayh-Dole and face challenges from universities that seek to retain their patent rights, but software patent disclosures alone are often limited and over-broad, and fail to enhance social welfare.[60] Requiring fuller disclosure of code generated on the NRC can therefore decrease the risk of over-patenting and increase AI innovation. The growth of the robust open-source and open science movements also suggests that an open-sourcing requirement for the NRC would not be a complete barrier to NRC usage.[61]

A strong argument for mandating open-sourcing also comes from the increasing private- sector reliance on trade secrets for IP protection in AI.[62] Some argue that this heightened emphasis on trade secret protection constitutes "artificial stupidity,"[63] as it has stifled innovation in AI by preventing disclosure, providing protection for a potentially unlimited duration, and attaching immediately and broadly to any output with perceivable economic value.[64] The reliance on secrecy, therefore, contravenes many of the principles described above—which argue that sharing code and data is crucial in AI—and results in significant AI industry consolidation and suboptimal levels of AI innovation.[65] This harkens back to the goal of the NRC discussed in Chapter 1: To address problems with AI research being concentrated in the hands of a few private-sector players. Because the NRC should explicitly avoid replicating these private-sector challenges, this lends additional support to a recommendation that the NRC should contemplate requiring researchers to share their research outputs.

In sum, while AI raises a host of novel IP issues (e.g., whether AI output is itself eligible for IP protection), we think the government can steer clear of many of these complications by tracking Bayh-Dole and the Uniform Guidance. The government should also consider conditions for requiring NRC researchers to disclose or share their research outputs under an open-access license.

Conclusion

As we have articulated in this White Paper, the ambitious call for an NRC has transformative potential for the AI research landscape.

Its biggest promise is to ensure more equitable access to core ingredients for AI research: compute and data. Leveling this playing field could shift the current ecosystem from one that focuses on narrow commercial problems to one that fosters basic, noncommercial AI research to ensure long-term national competitiveness, to solve some of the most pressing problems, and to rigorously interrogate AI models.

As we have spelled out in this White Paper, the NRC does raise a host of policy, legal, and normative questions. How can such compute resources be provided in a way that is expeditious and user-friendly, but does not preclude the potential cost savings from a publicly owned resource? How can the NRC be designed to adhere to the Privacy Act of 1974, which was animated by concerns about a national system of records that surveils its citizens? How can we ensure that NRC mitigates, rather than heightens, concerns about the unethical use of AI? And how can one prevent the NRC from becoming the biggest target for cyberattacks?

These are tough questions, and we hope to have sketched out our initial attempt at answers above. We are hopeful, if designed well, the NRC could help to realign the AI innovation space from one that is fixated with short-term private profit to one that is infused with long-term public values.

Glossary of Acronyms

ADP	Alberta Data Partnerships
ADR UK	Administrative Data Research UK
ADRF	Administrative Data Research Facility
AI	artificial intelligence
API	application programming interface
ARPA	Advanced Research Projects Agency
ARPANET	Advanced Research Projects Agency Network
ATO	authority-to-operate
ATO	Authorization to Operate
AWS	Amazon Web Services
CaaS	Compute as a Service
CIPSEA	Confidential Information Protection and Statistical Efficiency Act
CMS	Centers for Medicare & Medicaid Services
CPL	California Policy Lab
CPU	central processing unit
DFARS	Defense Federal Acquisition Regulation Supplement
DHS	U.S. Department of Homeland Security
DOD	U.S. Department of Defense
DOE	U.S. Department of Energy
DOT	U.S. Department of Transportation
DUA	Data Use Agreement
EBPA	Foundations for Evidence Based Policymaking Act or Evidence Act
EO	Executive Order
ESRC	Economic and Social Research Council
EULA	End-User Licensing Agreement
FAA	Federal Aviation Administration
FAR	Federal Acquisition Regulation
FDS	Federal Data Strategy
FedRAMP	Federal Risk and Authorization Management Program
FFRDC	Federally-Funded Research and Development Center
FIPS	Federal Information Processing Standards
FISMA	Federal Information Security Modernization Act
FSRDC	Federal Statistical Research Data Center
GAO	U.S. Government Accountability Office
GCP	Google Cloud Platform
GDPR	General Data Protection Regulation
GPS	Global Positioning System
GPU	graphics processing unit
GSA	U.S. General Services Administration
HAI	Stanford Institute for Human-Centered Artificial Intelligence
HECToR	High-End Computing Terascale Resource
HHS	U.S. Department of Health and Human Services
HIPAA	Health Insurance Portability and Accountability Act
HPC	high-performance computing
HTTP	Hypertext Transfer Protocol
HTTPS	Hypertext Transfer Protocol Secure
IC	U.S. Intelligence Community
IDA	Institute for Defense Analyses
IPTO	Information Processing Techniques Office
IRB	Institutional Review Board
JV	joint venture
LIDAR	Light Detection and Ranging
ML	machine learning
MOU	memorandum of understanding

NAIRR	National Artificial Intelligence Research Resource Task Force Act
NASA	National Aeronautics and Space Administration
NDAA	National Defense Authorization Act
NIH	National Institutes of Health
NISE	NSF's Directorate for Computer and Information Science and Engineering
NIST	National Institute of Standards and Technology
NIST SP	NIST Special Publications
NOAA	National Oceanic and Atmospheric Administration
NORC	National Opinion Research Center
NRC	National Research Cloud
NSCAI	National Security Commission on Artificial Intelligence
NSDS	National Secure Data Service
NSF	National Science Foundation
ODNI	Office of the Director of National Intelligence
OECD	Organization for Economic Co-operation and Development
OMB	U.S. Office of Management and Budget
ONS	Office for National Statistics
ORNL	Oak Ridge National Laboratory
OLCF	Oak Ridge Leadership Computing Facility
OSI	Open Systems Interconnection
OT	Other Transaction
PCLOB	Privacy and Civil Liberties Oversight Board
PHI	protected health information
PHS	Stanford Center for Population Health Sciences
PI	principal investigator
PII	personally identifiable information
PPP	public-private partnership
R&D	research and development
RFI	Request for Information
RFP	Request for Proposal
RIST	Research Organization for Information Science and Technology
SDSC	San Diego Supercomputer Center
SRCC	Stanford Research Computing Center
SSL	Secure Sockets Layer
STPI	Science & Technology Policy Institute
TLS	Transport Layer Security
UC Berkeley	University of California, Berkeley
UC San Diego	University of California, San Diego
UCLA	University of California, Los Angeles
USDA	U.S. Department of Agriculture
VRDC	CMS' Virtual Research Data Center
WIPO	World Intellectual Property Organization

Appendix

A. COMPUTING INFRASTRUCTURE COST COMPARISONS

This Appendix provides a sample cost-estimate comparison between a commercial cloud service, AWS, and a dedicated government HPC system, Summit. In sum, our estimations show that AWS P3 instances with comparable hardware to Summit would be 7.5 times as expensive as estimated costs under constant usage, and 2.8 times Summit's estimated costs under fluctuating demand.

Table 3 lists the three infrastructure models used in this comparison. Summit was used as the reference government HPC system because it is one of the DOE's newest systems and has hardware well-suited for AI research.[1] The other infrastructure model used is AWS EC2 P3.[2] Both are commonly used in AI research and general HPC applications. Other commercial cloud platforms, such as GCP or Azure, could also feasibly provide the infrastructure for the NRC. AWS EC2 P3 was used here because AWS has a robust cost calculator that allows for variable workloads.

The number of AWS instances were set such that those models would have the exact same number of GPUs as Summit. GPUs were the fixed variable because GPUs are the most important hardware for AI research applications, specifically deep learning. Both Summit and AWS P3 instances use NVIDIA V100 GPUs.

We conduct our cost comparison for the two infrastructure models over five years, as Summit's initial RFP documents include a five-year maintenance contract. AWS, however, only provides one-year or three-year pricing plans, so we extrapolated the five-year cost based on its three-year plan.

For the cost estimate of Summit, we based our calculation on the budget details in the original Department of Energy (DOE) Request for Proposal (RFP) in January 2014.[3] The RFP includes a $155 million maximum budget for building Summit, an expected $15 million maximum for the non-recurring engineering cost,[4] and around $15 million for five-year maintenance,[5] plus interest based on the U.S. Treasury securities at five-year constant maturity as specified in the price schedule.[6] Upon calculation, we estimated Summit costs around $192 million in total, which is consistent with public reporting of the cost of Summit.[7]

For the cost estimate of AWS, we used the AWS pricing calculator, choosing U.S. East (N. Virginia) as the data center and publicly available rates under the cheapest possible pricing plan (EC2 Instance Savings Plans). To approximate a negotiated discount, we applied a 10 percent discount based on the negotiated rate of one major university.

Since commercial cloud platform costs scale with how many instances are actually in use, two costs were calculated for each AWS model representing usage extremes: (1) with the infrastructure under constant usage; (2) with the infrastructure under dramatically fluctuating usage each day. For the daily spike traffic calculation, we set the model to run five days a week with 8.4 hours each day at peak performance. The maximum number of instances used is the same as one would use for constant use while the minimum number is zero. This workload setting is based on the assumption that GPUs used for training AI models sit idle 30 percent of the time.[8] These estimates should provide hard upper and lower bounds on costs for using each instance type.

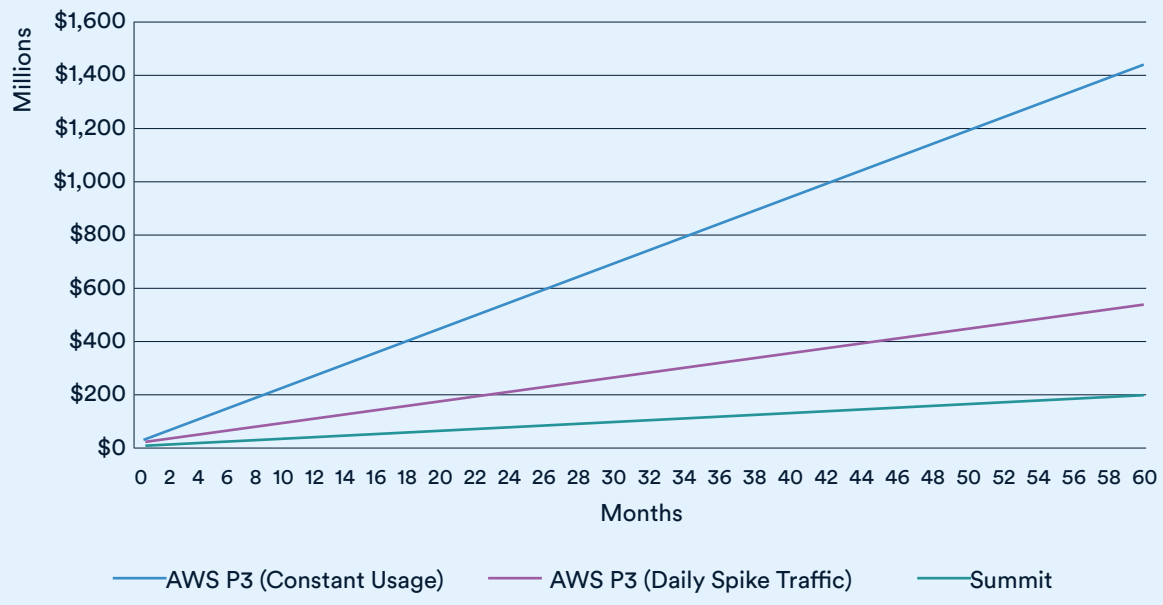
Figure 1 plots costs on the *y*-axis over a five-year period on the *x*-axis. The turquoise line indicates the cost to a Summit-like system and the purple and blue lines indicate the cost of the same AWS instances under variable and constant usage. Overall, this simple analysis corroborates the analysis conducted by Compute Canada, which found that commercial cloud "ranged from 4x to 10x more than the cost of owning and operating our own clusters."[9] Over five years and under constant usage, AWS P3 instances with comparable hardware to Summit would be 7.5 times as expensive as estimated costs. Under fluctuating demand, AWS P3 instances would cost 2.8 times Summit's estimated costs.

We note that this simple analysis omits many potential factors (see discussion in Chapter 2), but provides a starting point to understanding the considerable cost implications for the make-or-buy decision.

TABLE 3: SUMMIT & AWS COMPARISON

	GPUs	RAM	Network Bandwidth
Summit IBM AC922	27,648 (NVIDIA Volta V100)	2.8 PB	200 Gb/s
AWS p3dn.24xlarge (3456 nodes)	27,648 (NVIDIA Volta V100)	2.6 PB	100 Gb/s

FIGURE 1: ESTIMATED COST OF AWS INSTANCES COMPARED TO SUMMIT OVER 3 YEARS



B. FACILITATING PRIVATE DATASET SHARING

Unique IP challenges arise if researchers are permitted to share their own private datasets with the NRC. Indeed, researchers who "upload" proprietary data may be concerned about how other NRC users utilize that data.[10] Through interviews conducted for this White Paper, corporate stakeholders representing the entertainment industry, as well as other creative industries, have further expressed fear that researchers may upload and share data to which they do not hold rights. However, if the NRC does decide to facilitate private data-sharing, it should consider adopting two requirements to address these concerns: (1) The NRC should require all users to affirm they either have the original IP rights to the data or the data is already in the public domain; and (2) The NRC should have a scheme for its users to license their data.

(a) NRC users must own IP rights to the data they are uploading

Researchers uploading data need to agree that they own the intellectual property rights to the data prior to upload, or that the data is already in the public domain. This should be the case whether researchers share the data broadly with other researchers or simply use their data for their own private use.

Of course, despite mandating that uploaders guarantee legitimate ownership or public domain status of their uploaded IP, uploaders may nevertheless upload data they don't own the IP rights to. This may happen because computer engineers and researchers are not informed about IP law, anticipate that fair use will excuse their behavior, or simply hope not to get caught.[11] Industry stakeholders were also concerned that AI researchers would pull out "facts" from a copyrighted work (e.g., certain melodies in the chorus of a song) or apply certain algorithms to the work and "wrongly" claim a copyright over the transformed work. Whatever the case may be, this assembly of protected input data represents the "clearest copyright liability in the machine learning process" because assembling protected data violates the right to reproduction, and any preprocessing on the data could violate the right to derivative works.[12]

In interviews, corporate stakeholders expressed a desire to stymie the upload of copyrighted works by having the NRC itself assess whether uploaded data is already protected by copyright. Diligencing data can be completed manually, or by using such automated systems as *Content ID*, which is also used by corporations such as YouTube.[13] The former option would be very labor intensive,[14] whereas the latter may be prohibitively expensive,[15] so the value of addressing these concerns must be weighed against these burdensome costs.

Finally, it is unclear the extent to which uploading and sharing copyrighted data for machine learning amounts to fair use.[16] The most analogous case is Author's Guild v. Google Books.[17] In that case, Google scanned over 20 million books, many of which were copyright-protected, and assembled a corpus of machine-readable texts to power its Google Books service.[18] The 2nd Circuit held that Google Books' unauthorized reproductions of copyrighted works was transformative fair use, largely because Google Books provided information *about* books through small snippets, without threatening the rights-holders' core protectable expression in the books.[19] While some have opined that the Author's Guild holding categorically protects using copyrighted material in datasets for machine learning purposes,[20] many legal scholars are not so sure about such a broad holding, especially because fair use is so fact-intensive.[21] Indeed, while Google Books used copyrighted works for a nonexpressive purpose, Sobel notes that machine learning models may increasingly be able to glean value from a work's expressive aspects.[22] Therefore, until courts and legislators provide more clarity on the applicability of fair use in the machine learning context, the NRC should still require data uploaders to attest that they own the rights to the data.

(b) Users must be able to license their data to other users.

If the NRC enabled private data sharing, users would need to make clear what rights other NRC users have over the uploaders' shared data. The NRC would have two basic options for creating IP licensing schemes: (1) The NRC

could permit researchers to use whatever IP license they wish when sharing their private data; or (2) The NRC could mandate a uniform license across the board for all data that is uploaded.

(1) Researcher's Choice of License

Allowing researchers to craft their own IP licensing agreements when sharing private data with other researchers would be the most frictionless solution from the perspective of the uploader; it would allow them to share exactly what they want and restrict use to only certain contexts. This choice of license seems to be important to data sharers.[23] Indeed, many data scientists and engineers have written guides advising members of the open-source community on how they should go about choosing specific licenses for their work.[24] GitHub, an open-source code-sharing platform, permits its users to choose from dozens of licenses,[25] and FigShare, a data-sharing platform for researchers, likewise supports a host of different Creative Commons licenses.[26] Some datasets even have their own custom IP licensing agreements. The Twitter academic dataset, for instance, is licensed according to Twitter's own developer agreement and noncommercial use policies, not to an existing open-source license.[27]

However, there are disadvantages to such flexibility. Just because different licenses might be allowed doesn't mean these licenses will be fully understood by all users. Adopting multiple licenses may result in increased accidental infringement. Indeed, a study conducted by the Institute of Electrical and Electronics Engineers found that "although [software] developers clearly understood cases involving one license, they struggled when multiple licenses were involved,"[28] and in particular, were found to "lack the knowledge and understanding to tease apart license interactions across multiple situations."[29]

In particular, researchers unfamiliar with the allowances provided by different data licenses, in contexts where more than one license is implemented, may lead to certain licenses being violated. For example, when researchers were surveyed regarding their understanding of copyright transfer agreements in the IP commercialization process, they only demonstrated an average 33 percent score on a knowledge-testing survey.[30]

(2) Uniform Licensing Agreement

The second option available to the NRC would be to mandate that all private data be licensed under a single uniform license. For the NRC administration itself, this may be the more straightforward option, since users could be notified upon login about the appropriate use of data. The disadvantage of this strategy is that it may deter would-be researchers who would share data under a narrower license.[31] Given the desire to allow researchers to innovate freely, there may be concerns about adopting a restrictive licensing agreement. Nonetheless, several options of licensing agreements would still be available for adoption, and this pathway would require choosing a uniform agreement from these options, with the possibility of allowing an opt-out of this default license.

If the NRC were to implement a uniform license, it could look to the licensing agreements leveraged by institutional research clouds, such as the Harvard Dataverse as an analogy in determining best practices for its own licensing agreements. The model adopted by the Dataverse is a default use of the CC0 Public Domain Dedication "because of its name recognition in the scientific community" and its "use by repositories as well as scientific journals that require the deposit of open data."[32] Like an unrestricted Creative Commons or Open Data license, a public domain license would allow the data it governs to be used in any context, even commercial ones, and would also allow reproduction and creation of derivatives from the data.

Alternatively, the NRC could have a default open license while also permitting researchers to choose from a handful of more restrictive licenses if they wish. For example, the Harvard Dataverse notably allows uploaders to optout of the CC0 if needed and specify custom terms of use. The Australian Research Data Commons and data-sharing platform FigShare[33] also use a default CC0 license but nevertheless permit researchers to use a conditioned Creative Commons license. These conditioned licenses can, for instance, require attribution to the

original owner, prevent exact reproduction, or only allow use for noncommercial contexts. This may also help accommodate researchers who seek to upload datasets incorporating third-party data that holds a more restrictive license, since a "combined dataset will adopt the most restrictive condition(s) of its component parts."[34]

If the NRC goes down this route of giving users the choice of a narrower license, it would also shift some liability to users—or to the NRC itself—by relying on users to abide by the license. Approaches to enforcement would vary, depending on the amount of responsibility in enforcement, and by extension liability the NRC seeks to take on. For example, in the Harvard Dataverse, if an uploader decides to opt out of a default open license and pursue their own custom licensing agreement over uploaded data, the Dataverse's General Terms of Use absolve this particular cloud from resource-heavy enforcement responsibilities by stating that it "has no obligation to aid or support either party of the Agreement in the execution or enforcement of the Data Use Agreement's terms."[35]

C. CURRENT STATE OF AI ETHICS FRAMEWORKS

AI ethics frameworks (or principles, guidelines) attempt to address the ethical concerns related to the development, deployment, and use of AI within prospective organizations. We briefly discuss the current landscape of AI ethics frameworks, while noting that this is still an emergent topic without broad consensus.

Between 2015 and 2020, governments, technology companies, international organizations, professional organizations, and researchers around the world have published some 117 documents related to AI ethics.[36] These frameworks aim to tackle the disruptive potential of AI technologies by producing normative principles and "best practice" recommendations.[37] Due to the prominence of essentially contested concepts in AI ethics—i.e., words such as fairness, equity, privacy that have different meanings for different audiences[38]—as well as the lack of binding professional history and accountability mechanisms, those frameworks are often high level and self-regulatory, posing little threat to potential breaches to ethical conduct.[39]

Federal Frameworks

In the United States, there is no central guiding framework on the responsible development and application of AI across the federal government. Some government agencies have adopted or are in the process of adopting their own AI framework, while others have not published such guidelines. The following are published federal AI ethical frameworks as of August 2021:

- After 15 months of deliberation with leading AI experts, the Department of Defense (DOD) adopted a series of ethical principles for the use of AI in February 2020 that align with the existing DOD mission and stakeholders.[40]

- The General Services Administration (GSA), tasked by the Office of Management and Budget (OMB) in the Federal Data Strategy 2020 Action Plan, developed a Data Ethics Framework in February 2020 to help federal personnel make ethical decisions as they acquire, manage, and use data.[41]

- The Government Accountability Office (GAO) developed an AI accountability framework in June 2020 for federal agencies and other entities involved in the design, development, deployment, and continuous monitoring of AI systems to help ensure accountability and responsible use of AI.[42]

- The Office of the Director of National Intelligence (ODNI) released the *Principles of AI Ethics for the Intelligence Community* in July 2020 to guide the intelligence community's (IC) ethical development and use of AI to solve intelligence problems.[43]

- The National Security Commission on Artificial Intelligence (NSCAI) published a set of best practices in July 2020 (later revised and

integrated into the Commission's 2021 Final Report) for agencies critical to national security to implement as a paradigm for the responsible development and fielding of AI systems.[44]

While these frameworks can help guide the NRC's approach to ethics, we refrain from recommending a specific framework for several reasons. First, despite growing calls for applied ethics in the AI community, developing an AI ethics framework is still an emerging area. The lack of a unified government standard poses challenges to the establishment of the NRC's ethics review process.

Second, there are, in fact, significant differences among ethics frameworks published by various federal agencies. For example, NSCAI laid out differences between its recommended practices and those by DOD and IC.[45] Moreover, among the five frameworks above, the GSA Framework focused only on the ethical conduct of federal employees when dealing with data while others focused on the ethical development and application of AI systems specifically.

Third, the ethics framework for adopting AI technology may be different from a framework for assessing research. Most federal agencies develop frameworks to guide the use of AI-driven solutions for agency-specific tasks. For example, DOD's ethical principles only apply to defense-specific combat or noncombat AI systems.[46] In the absence of a central federal guideline, the NRC should not adopt a framework by a particular agency because these frameworks are not necessarily designed for the wide range of research contemplated for the NRC. The work on frameworks may nonetheless provide a useful starting point for NRC's ethics process.

D. STAFFING AND EXPERTISE

As noted throughout this White Paper, the success of the NRC will depend on human resources—both within the NRC as well as across government— to resolve the many challenges the NRC promises to tackle. While we refrain from providing an organizational chart, we list the dimensions where staffing and expertise will be critical to the success of the NRC. This list is not meant to be exhaustive, but to highlight the vital importance of human resources.

Human Resource Areas

- Computing
    - System administrators
    - Data center engineers
    - Research software engineers
    - Research application developers
- Data
    - Data officers
    - Agency liaisons
    - Data architects
    - Data scientists
- Grant administrators
- Contracting officers
- Support and training staff
- Privacy staff (technical and legal)
- Ethics staff
- Cybersecurity staff

Endnotes

Executive Summary

1 Klaus Schwab, The Fourth Industrial Revolution (2016).
2 Tae Yano & Moonyoung Kang, *Taking Advantage of Wikipedia in Natural Language Processing*, Carnegie Mellon U. (2008), https://www.cs.cmu.edu/~taey/pub/wiki.pdf.
3 *See, e.g.,* Anthony Alford, *Google Trains Two Billion Parameter AI Vision Model,* InfoQ (June 22, 2021), https://www.infoq.com/news/2021/06/google-vision-transformer/; Anthony Alford, *OpenAI Announces GPT-3 AI Language Model with 175 Billion Parameters,* InfoQ (June 2, 2020), https://www.infoq.com/news/2020/06/openai-gpt3-language-model/.
4 *AlphaGo*, DeepMind (2021), https://deepmind.com/research/case-studies/alphago-the-story-so-far/.
5 Benjamin F. Jones & Lawrence H. Summers, *A Calculation of the Social Returns to Innovation* (Nat'l Bureau of Econ. Research, Working Paper No. 27863, 2020); J.G. Tewksbury, M.S. Crandall & W.E. Crane, *Measuring the Societal Benefits of Innovation*, 209 Sci. Mag. 658-62 (1980); *see also* National Academies of Sciences, Engineering, and Medicine, Returns to Federal Investments in the Innovation System (2017)
6 Stuart Zweben & Betsy Bizot, 2019 Taulbee Survey: Total Undergrad CS Enrollment Rises Again, but with Fewer New Majors; Doctoral Degree Production Recovers From Last Year's Dip (2019).
7 Jathan Sadowski, *When Data is Capital: Datafication, Accumulation, and Extraction*, 2019 Big Data & Soc'y 1 (2019).
8 Amy O'Hara & Carla Medalia, *Data Sharing in the Federal Statistical System: Impediments and Possibilities*, 675 Annals Am. Acad. Pol. & Soc. Sci. 138, 140-41 (2018).
9 Nat'l Security Comm'n on Artificial Intelligence, Final Report 186 (2021).
10 Stan. U. Inst. for Human-Centered Artificial Intelligence, 2021 Artificial Intelligence Index Report 118 (2021).
11 *Id.*
12 *Id.*
13 Neil C. Thompson, Shuning Ge & Yash M. Sherry, *Building the Algorithm Commons: Who Discovered the Algorithms that Underpin Computing in the Modern Enterprise?*, 11 Global Strategy J. 17-33 (2020).
14 *See, e.g.*, U.S. Gov't Accountability Office, Federal Agencies Need to Address Aging Legacy Systems (2016); U.S. Gov't Accountability Office, Cloud Computing: Agencies Have Increased Usage and Realized Benefits, but Cost and Savings Data Need To Be Better Tracked (2019).
15 David Freeman Engstrom, Daniel E. Ho, Catherine M. Sharkey & Mariano-Florentino Cuéllar, Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies 6, 71-72 (2020).
16 William M. (Mac) Thornberry National Defense Authorization Act for Fiscal Year 2021, Pub. L. No. 116-283, § 5106.
17 *The Biden Administration Launches the National Artificial Intelligence Research Resource Task Force*, The White House (June 10, 2021), https://www.whitehouse.gov/ostp/news-updates/2021/06/10/the-biden-administration-launches-the-national-artificial-intelligence-research-resource-task-force/.
18 William M. (Mac) Thornberry National Defense Authorization Act for Fiscal Year 2021, Pub. L. No. 116-283, § 5107 (g).
19 Nat'l Security Comm'n on Artificial Intelligence, *supra* note 9, at 191.
20 *See, e.g.*, *Cloudbank*, https://www.cloudbank.org; *Fact Sheet: National Secure Data Service Act Advances Responsible Data Sharing in Government*, Data Coalition (May 13, 2021), https://www.datacoalition.org/fact-sheet-national-secure-data-service-act-advances-responsible-data-sharing-in-government/.
21 Steve Lohr, *Universities and Tech Giants Back National Cloud Computing Project*, N.Y. Times (June 30, 2020), https://www.nytimes.com/2020/06/30/technology/national-cloud-computing-project.html; John Etchemendy & Fei-Fei Li, *National Research Cloud: Ensuring the Continuation of American Innovation*, Stan. U. Inst. for Human-Centered Artificial Intelligence, (Mar. 28, 2020), https://hai.stanford.edu/news/national-research-cloud-ensuring-continuation-american-innovation.
22 Jennifer Villa & Dave Troiano, *Choosing Your Deep Learning Infrastructure; The Cloud vs. On-Prem Debate*, Determined AI (July 30, 2020), https://determined.ai/blog/cloud-v-onprem/; *Is HPC Going to Cost Me a Fortune?*, InsideHPC (last visited July 23, 2021), https://insidehpc.com/hpc-basic-training/is-hpc-going-to-cost-me-a-fortune/.
23 *See, e.g., US Plans $1.8 Billion Spend on DOE Exascale Supercomputing*, HPCwire (Apr. 11, 2018), https://www.hpcwire.com/2018/04/11/us-plans-1-8-billion-spend-on-doe-exascale-supercomputing/; *Federal Government*, Advanced HPC (last visited July 23, 2021), https://www.advancedhpc.com/pages/federal-government; *United States Continues to Lead World In Supercomputing*, U.S. Dep't. Energy (Nov. 18, 2019), https://www.energy.gov/articles/united-states-continues-lead-world-supercomputing.
24 *See NSF Funds Five New XSEDE-Allocated Systems,* Nat'l Sci. Found. (Aug. 10, 2020), https://www.xsede.org/-/nsf-funds-five-new-xsede-allocated-systems.
25 *Cloudbank*, *supra* note 20.
26 *See, e.g., National Data Service*, http://www.nationaldataservice.org; *The Open Science Data Cloud*, https://www.opensciencedatacloud.org; *Harvard Dataverse*, https://dataverse.harvard.edu; *FigShare*, https://figshare.com.
27 *FedRAMP*, https://www.fedramp.gov.
28 *See Fact Sheet: National Secure Data Service Act Advances Responsible Data Sharing in Government,* Data Coalition (May 13, 2021), https://www.datacoalition.org/fact-sheet-national-secure-data-service-act-advances-responsible-data-sharing-in-government/.
29 *See Administrative Data Research Facility*, Coleridge Initiative, https://coleridgeinitiative.org/adrf/ (last visited July 26, 2021).
30 *See Landsat Data Access*, U.S. Geological Survey, https://www.usgs.gov/core-science-systems/nli/landsat/landsat-data-access (last visited July 23, 2021); Fed. Geographic Data Comm., The Value Proposition for Landsat Applications (2014); Crista L. Straub, Stephen R. Koontz & John B. Loomis, Economic Valuation of Landsat Imagery (2019).

31 *See* Bipartisan Pol'y Ctr., Barriers to Using Government Data: Extended Analysis of the U.S. Commission on Evidence-Based Policymaking's Survey of Federal Agencies and Offices 18-20 (2018); *see also* U.S. Dep't of Health & Human Services, The State of Data Sharing at the U.S. Department of Health and Human Services 4 (2018) (describing how data at the agency is "largely kept in silos with a lack of organizational awareness of what data are collected across the Department and how to request access.").
32 Privacy Act, 5 U.S.C. § 552a (1974).
33 Michael S. Bernstein et al., *ESR: Ethics and Society Review of Artificial Intelligence Research*, Cornell. U. (July 9, 2021), https://arxiv.org/pdf/2106.11521.pdf.
34 Courtenay R. Bruce et al., *An Embedded Model for Ethics Consultation: Characteristics, Outcomes, and Challenges*, 5 AJOB Empirical Bioethics 8 (2014).

Introduction

1 *National Research Cloud Call to Action*, Stan. U. Inst. For Human-Centered Artificial Intelligence, https://hai.stanford.edu/national-research-cloud-joint-letter.
2 *See id.*; John Etchemendy & Fei-Fei Li, *National Research Cloud: Ensuring the Continuation of American Innovation*, Stan. U. Inst. For Human-Centered Artificial Intelligence (Mar. 28, 2020), https://hai.stanford.edu/news/national-research-cloud-ensuring-continuation-american-innovation.
3 William M. (Mac) Thornberry National Defense Authorization Act for Fiscal Year 2021, Pub. L. No. 116-283, § 5106.
4 *The Biden Administration Launches the National Artificial Intelligence Research Resource Task Force*, The White House (June 10, 2021), https://www.whitehouse.gov/ostp/news-updates/2021/06/10/the-biden-administration-launches-the-national-artificial-intelligence-research-resource-task-force/.
5 Privacy Act of 1974, 5 U.S.C. § 552a (2012).
6 Foundations for Evidence-Based Policymaking Act of 2017, Pub. L. No. 115-435, 132 Stat. 5529 (2019).
7 *See Fact Sheet: National Secure Data Service Act Advances Responsible Data Sharing in Government,* Data Coalition (May 13, 2021), https://www.data-coalition.org/fact-sheet-national-secure-data-service-act-advances-responsible-data-sharing-in-government/.
8 *See, e.g.,* Facial Recognition and Biometric Technology Moratorium Act, S. 4084, 116th Cong. (2020); Bhaskar Chakravorti, *Biden's 'Antitrust Revolution' Overlooks AI—at Americans' Peril*, Wired (July 27, 2021), https://www.wired.com/story/opinion-bidens-antitrust-revolution-overlooks-ai-at-americans-peril/.

Chapter 1

1 *See* Stephen Breyer, Regulation and Its Reform (1982); Clifford Winston, Government Failure Versus Market Failure (2006).
2 *Largest Companies by Market Cap*, Companies Market Cap (2021), https://companiesmarketcap.com.
3 Stan. U. Inst. for Human-Centered Artificial Intelligence, 2021 Artificial Intelligence Index Report 93 (2021).
4 *See, e.g.,* Mary L. Gray & Siddarth Suri, Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass (2019); Craig Webster & Stanislav Ivanov, Robotics, Artificial Intelligence, and the Evolving Nature of Work 132-35 (2019); Weiyu Wang & Keng Siau, *Artificial Intelligence, Machine Learning, Automation, Robotics, Future of Work and Future of Humanity: A Review and Research Agenda*, 30 J. Database Mgmt. 61 (2019).
5 *AlphaFold: A Solution to a 50-Year-Old Grand Challenge in Biology*, DeepMind (Nov. 30, 2020), https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology.
6 Tanha Talaviya et al., *Implementation of Artificial Intelligence in Agriculture for Optimisation of Irrigation and Application of Pesticides and Herbicides*, 4 Artificial Intelligence in Agriculture 58 (2020).
7 Greg Allen & Taniel Chan, *Artificial Intelligence and National Security,* Harv. Kennedy Sch. Belfer Ctr. (July 2017), https://www.belfercenter.org/publication/artificial-intelligence-and-national-security.
8 Stan. U. Inst. for Human-Centered Artificial Intelligence, *supra* note 3.
9 Jeffrey Ding, Deciphering China's AI Dream (2018).
10 Fugaku is being used extensively for AI research initiatives. *See* Atsushi Nukariya et al., *HPC and AI Initiatives for Supercomputer Fugaku and Future Prospects*, Fujitsu (Nov. 11, 2020), https://www.fujitsu.com/global/about/resources/publications/technicalreview/2020-03/article09.html.
11 Eng'g & Physical Sciences Research Council, The Impact of HECToR (2014).
12 Joshua New, Why the United States Needs a National Artificial Intelligence Strategy and What it Should Look Like (2018).
13 Maggie Miller, *White House Establishes National Artificial Intelligence Office*, The Hill (Jan. 12, 2021), https://thehill.com/policy/cybersecurity/533922-white-house-establishes-national-artificial-intelligence-office.
14 *See* Fast Track Action Comm. on Strategic Computing, National Strategic Computing Initiative Update: Pioneering the Future of Computing (2019), https://www.nitrd.gov/pubs/National-Strategic-Computing-Initiative-Update-2019.pdf.
15 Nat'l Security Comm'n on Artificial Intelligence, Final Report (2021).
16 *The COVID-19 High Performance Computing Consortium*, COVID-19 HPC Consortium, https://covid19-hpc-consortium.org.
17 *See* Aaron L. Friedberg, *Science, the Cold War, and the American State*, 20 Diplomatic Hist. 107, 112 (1996); Sean Pool & Jennifer Erickson, *The High Return on Investment for Publicly Funded Research*, Ctr. for Am. Progress (Dec. 10, 2012), https://www.americanprogress.org/issues/economy/reports/2012/12/10/47481/the-high-return-on-investment-for-publicly-funded-research/.
18 Peter L. Singer, Federally Supported Innovations: 22 Examples of Major Technology Advances That Stem from Federal Research Support 14-15 (2014).
19 Nat'l Research Council, Government Support for Computing Research 136-55 (1999).
20 Nat'l Security Comm'n on Artificial Intelligence, *supra* note 15, at 185.
21 Philippe Aghion, Benjamin F. Jones & Charles I. Jones, *Artificial Intelligence and Economic Growth*, *in* The Economics of Artificial Intelligence: An Agenda 237 (2019).

22 Ian Moll, *The Myth of the Fourth Industrial Revolution*, 68 Theoria 1 (2021); *see also* Tim Unwin, *5 Problems with 4th Industrial Revolution*, ICT Works (Mar. 23, 2019), https://www.ictworks.org/problems-fourth-industrial-revolution/.
23 *See, e.g.*, Geoffrey A. Manne & Joshua D. Wright, *Google and the Limits of Antitrust: The Case Against the Antitrust Case Against Google*, 34 Harv. J.L. & Pub. Pol'y 1 (2011); Lina M. Khan, *Amazon's Antitrust Paradox*, 126 Yale L.J. 710 (2016).
24 David Patterson et al., *Carbon Emissions and Large Neural Network Training*, Cornell U. (Apr. 23, 2021), https://arxiv.org/pdf/2104.10350.pdf. To be clear, however, the study found that training other sophisticated but smaller NLP models such as Meena and T5 required approximately 96 and 48 tons of carbon dioxide, respectively. *Id.* Another study found that the training state-of-the-art NLP models produced approximately 626,000 pounds (313 tons) of carbon dioxide, five times the lifetime emissions of the average car in the United States. Emma Strubell, Ananya Ganesh & Andrew McCallum, *Energy and Policy Considerations for Deep Learning in NLP*, Cornell U. (2019), https://arxiv.org/pdf/1906.02243.pdf.
25 *Calculate Your Carbon Footprint*, The Nature Conservancy, https://www.nature.org/en-us/get-involved/how-to-help/carbon-footprint-calculator/.
26 Economic studies in other fields also show that increasing access, supply, or quality of certain goods without appropriate pricing mechanisms or regulatory interventions can lead to over-use and waste. *See, e.g.*, Chengri Ding & Shunfeng Song, *Traffic Paradoxes and Economic Solutions*, 1 J. Urban Mgmt. 63 (2012) (roads and traffic congestion); Ari Mwachofi & Assaf F. Al-Assaf, *Health Care Market Deviations from the Ideal Market*, 11 Sultan Qaboos Univ. Med. J. 328 (2011) (doctors and quality of care).
27 *See* Emily M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, 2021 Proceedings ACM Conf. on Fairness, Accountability & Transparency 610 (2021).
28 *See* Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 Proceeding Machine Learning Res. 1 (2018); Inioluwa Deborah Raji & Joy Buolamwini, *Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products*, 2019 Proceedings AAAI/ACM Conf. on AI, Ethics & Soc'y 429 (2019).
29 *See* Virginia Eubanks, Automating Inequality (2018); Cathy O'Neil, Weapons of Math Destruction (2016).
30 *See* Christopher Whyte, *Deepfake News: AI-Enabled Disinformation as a Multi-Level Public Policy Challenge*, 5 J. Cyber Pol'y 199 (2020); Jeffrey Dastin, *Amazon Scrapes Secret AI Recruiting Tool that Showed Bias Against Women*, Reuters (Oct. 10, 2018), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G; James Vincent, *Google 'Fixed' its Racist Algorithm by Removing Gorillas from its Image-Labeling Tech*, The Verge (Jan. 12, 2018), https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai;.
31 Kate Crawford, Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence 211 (2021) ("AI systems are built to see and intervene in the world in ways that primarily benefit the states, institutions, and corporations that they serve. In this sense, AI systems are expressions of power that emerge from wider economic and political forces, created to increase profits and centralize control for those who wield them.").
32 Elizabeth Gibney, *Self-Taught AI is Best Yet at Strategy Game Go*, Nature (Oct. 18, 2017), https://www.nature.com/news/self-taught-ai-is-best-yet-at-strategy-game-go-1.22858.
33 Bill Schackner, *Carnegie Mellon's Prestigious Computer Science School has a New Leader*, Pittsburgh Post-Gazette (Aug. 8, 2019), https://www.post-gazette.com/news/education/2019/08/08/Carnegie-Mellon-University-computer-science-Martial-Hebert-dean-artificial-intgelligence-google-robotics/stories/201908080096.
34 Bipartisan Pol'y Ctr, Cementing American Artificial Intelligence Leadership: AI Research & Development (2020).
35 Nur Ahmed & Muntasir Wahed, *The De-Democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research*, Cornell U. (Oct. 22, 2020), https://arxiv.org/pdf/2010.15581.pdf.
36 *Id.*
37 Fei-Fei Li, *America's Global Leadership in Human-Centered AI Can't Come From Industry Alone*, The Hill (July 6, 2021), https://thehill.com/opinion/technology/561638-americas-global-leadership-in-human-centered-ai-cant-come-from-industry?rl=1.
38 Cade Metz, *A.I. Researchers Are Making More Than $1 Million, Even at a Nonprofit*, N.Y. Times (Apr. 19, 2018), https://www.nytimes.com/2018/04/19/technology/artificial-intelligence-salaries-openai.html.
39 Stan. U. Inst. for Human-Centered Artificial Intelligence, *supra* note 3, at 118.
40 Michael Gofman & Zhao Jin, *Artificial Intelligence, Education, and Entrepreneurship*, SSRN (Sept. 17, 2019), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3449440.
41 Jathan Sadowski, *When Data is Capital: Datafication, Accumulation, and Extraction*, 2019 Big Data & Soc'y 1 (2019).
42 For example, researchers have clamored for Facebook to share some of its proprietary data so they can better understand the effect of social media on politics and societal discourse. Simon Hegelich, *Facebook Needs to Share More with Researchers*, Nature (Mar. 24, 2020), https://www.nature.com/articles/d41586-020-00828-5.
43 Ashlee Vance, *This Tech Bubble Is Different*, Bloomberg (Apr. 14, 2011), https://www.bloomberg.com/news/articles/2011-04-14/this-tech-bubble-is-different.
44 Amy O'Hara & Carla Medalia, *Data Sharing in the Federal Statistical System: Impediments and Possibilities*, 675 Annals Am. Acad. Pol. & Soc. Sci. 138, 140-41 (2018).
45 Dario Amodei & Danny Hernandez, *AI and Compute*, Open AI (May 16, 2018), https://openai.com/blog/ai-and-compute/.
46 *See, e.g.,* Ahmed & Wahed, *supra* note 35; Ian Sample, *'We Can't Compete': Why Universities Are Losing Their Best AI Scientists*, The Guardian (Nov. 1, 2017), https://www.theguardian.com/science/2017/nov/01/cant-compete-universities-losing-best-ai-scientists.
47 Neil C. Thompson, Shuning Ge & Yash M. Sherry, *Building the Algorithm Commons: Who Discovered the Algorithms that Underpin Computing in the Modern Enterprise?,* 11 Global Strategy J. 17-33 (2020).
48 Minkyung Baek, *RoseTTAFold: Accurate Protein Structure Prediction Accessible to All*, U. Wash. Inst. for Protein Design (July 15, 2021), https://www.ipd.uw.edu/2021/07/rosettafold-accurate-protein-structure-prediction-accessible-to-all/; Minkyung Baek et al., *Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network*, Sci. Mag. (July 15, 2021), https://science.sciencemag.org/content/sci/early/2021/07/19/science.abj8754.full.pdf.
49 *How Diplomacy Helped to End the Race to Sequence the Human Genome*, Nature (June 24, 2020), https://www.nature.com/articles/d41586-020-01849-w.
50 Joel Klinger et al., *A Narrowing of AI Research?*, Cornell U. (Nov. 17, 2020), https://arxiv.org/pdf/2009.10385.pdf.
51 *Id*.
52 Alex Tamkin et al., *Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models*, Cornell U. (Feb. 4, 2021), https://arxiv.org/pdf/2102.02503.pdf.

53 Those 5 were Boston, San Francisco, San Jose, Seattle and San Diego. *See* Robert D. Atkinson, Mark Muro & Jacob Whiton, *The Case for Growth Centers: How to Spread Tech Innovation Across America*, Brookings (Dec. 9, 2019), https://www.brookings.edu/research/growth-centers-how-to-spread-tech-innovation-across-america/.
54 Interview with Professor Erik Brynjolfsson, Director, Stanford Digital Economy Lab (2021).
55 Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 Cal. L. Rev. 671 (2016).

Chapter 2

1 "Principal Investigator" status may differ from university to university, but typically represents the core faculty that are eligible to oversee research projects at their home institutions.
2 *See* Beth Jensen, *AI Index Diversity Report: An Unmoving Needle*, Stan. U. Inst. for Human-Centered Artificial Intelligence (May 3, 2021), https://hai.stanford.edu/news/ai-index-diversity-report-unmoving-needle.
3 For a perspective, for instance, on the importance of modeling and simulation in physics, see Karen E. Wilcox, Omar Ghattas & Patrick Heimbach, *The Imperative of Physics-Based Modeling and Inverse Theory in Computational Science*, 1 Nature Comp. Sci. 166 (2021).
4 15 U.S.C. § 9415 (emphasis added).
5 *Id.* (emphasis added).
6 Contemporaneous accounts corroborate this core focus. The National Security Commission on AI, for instance, describes the proposal as "provid[ing] verified researchers and students subsidized access to scalable compute resources" with a specific reference to the "compute divide" that has left "middle- and lower-tier universities [lacking] the resources necessary for cutting-edge AI research." Nat'l Security Comm'n on A.I., Final Report 191, 197 (2021) (emphasis added). Upon the announcement of the NRC legislation, Jeff Dean, SVP of Google Research and Google Health, noted, "A National AI Research Resource will help accelerate US progress in artificial intelligence and advanced technologies by providing academic researchers access to the cloud computing resources necessary for experiments at scale." Brandi Vincent, C*ongress Inches Closer to Creating a National Cloud for AI Research*, NextGov (July 2, 2020), https://www.nextgov.com/emerging-tech/2020/07/congress-inches-closer-creating-national-cloud-ai-research/166624/ (emphasis added). Others have suggested that "researchers" under NRC could include individuals at small businesses, start-up companies, non-profits, and certain technology firms. One co-sponsor of the legislation, for instance, suggested that NRC resources should be provided to "developers" and "entrepreneurs." *Portman, Heinrich Introduce Bipartisan Legislation to Develop National Cloud Computer for AI Research*, Rob Portman, U.S. Senator for Ohio (June 4, 2020), https://www.portman.senate.gov/newsroom/press-releases/portman-heinrich-introduce-bipartisan-legislation-develop-national-cloud.
7 *Frequently Asked Questions About Small Businesses*, U.S. Small Bus. Admin. Office of Advoc. (Oct. 2020), https://cdn.advocacy.sba.gov/wp-content/uploads/2020/11/05122043/Small-Business-FAQ-2020.pdf.
8 Louise Balle, *Information on Small Business Startups*, Houston Chron., https://smallbusiness.chron.com/information-small-business-startups-2491.html.
9 Such entities could potentially collaborate with academic partners, and the NRC would of course also need to set rules about collaborator eligibility.
10 PI status provides a level of standardization across faculty compared to other metrics, such as tenure-track or designation as research faculty. For example, the University of Michigan appoints individuals focused on full-time research as "research faculty," which is not a tenure track position. In contrast, research faculty at Purdue are eligible for tenure-track. Distinct from the categorization used by both universities, MIT designates full-time researchers as "academic staff" rather than faculty. All three types of researchers, however, qualify for principal investigator status at their respective universities. Some universities go further by providing temporary PI status to non-PI status individuals affiliated with the university for a single project (including all three universities mentioned previously).
11 *Community & Education Resource Requests*, CloudBank, https://www.cloudbank.org/training/cloudbank-community#toc-eligibilit-36nfpcrS.
12 *Apply for an Account*, Compute Canada, https://www.computecanada.ca/research-portal/account-management/apply-for-an-account/.
13 Nat'l Sci. Bd., Science & Engineering Indicators 2016, Academic Research and Development 72 (2016).
14 *Id*.
15 *College Enrollment in the United States from 1965 to 2019 and Projections up to 2029 for Public and Private Colleges,* Statista (Jan. 2021), https://www.statista.com/statistics/183995/us-college-enrollment-and-projections-in-public-and-private-institutions/.
16 *Colaboratory – Frequently Asked Questions*, Google, https://research.google.com/colaboratory/faq.html.
17 *Weekly Maximum GPU Usage*, Kaggle (2019), https://www.kaggle.com/general/108481.
18 *Community & Education Resource Requests*, *supra* note 11.
19 *Merit Review: Why You Should Volunteer to Serve as an NSF Reviewer*, Nat'l Sci. Found., https://www.nsf.gov/bfa/dias/policy/merit_review/reviewer.jsp#1.
20 *See XSEDE Campus Champions,* XSEDE, https://www.xsede.org/community-engagement/campus-champions.
21 Compute Canada, for instance, provides access to 15% of PIs to increased compute capacity based on a merit competition. In 2021, Compute Canada completed its review of 650 research submissions in about five months with only 80 volunteer reviewers from Canadian academic institutions to assess the scientific merit of the proposal. *Resource Allocation Competitions*, Compute Canada, https://www.computecanada.ca/research-portal/accessing-resources/resource-allocation-competitions/; *2021 Resource Allocations Competition Results*, Compute Canada, https://www.computecanada.ca/research-portal/accessing-resources/resource-allocation-competitions/rac-2021-results/. Compare this with CloudBank, which allocates compute resources by leveraging NSF's grant administration process: In 2019, NSF needed 30,000 volunteer reviewers to handle over 40,000 proposals, with each proposal requiring about 10 months to process from start to finish. Nat'l. Sci. Found., Merit Review Process: Fiscal Year 2019 Digest (2020); *NSF Proposal and Award Process*, Nat'l Sci. Found., https://www.nsf.gov/attachments/116169/public/nsf_proposal_and_award_process.pdf.
22 Another boundary question will be the resource allocation to PIs that are affiliated both with universities and with private companies. As a default, NRC resources should go toward academic projects, and not subsidize work that is conducted in private researcher capacity.
23 *Resource Allocation Competitions*, *supra* note 21.
24 *Simplifying Cloud Services*, Sci. Node (Dec. 2, 2019), https://sciencenode.org/feature/An%20easier%20cloud.php.
25 *Frequently Asked Questions (FAQ)*, CloudBank, https://www.cloudbank.org/faq.
26 *Simplifying Cloud Services*, *supra* note 25.
27 *Id.*

28 *Id.*
29 *Frequently Asked Questions (FAQ)*, *supra* note 26.
30 *Frequently Asked Questions (FAQs) for Budgeting for Cloud Computing Resources via CloudBank in NSF Proposals*, Nat'l Sci. Found., https://www.nsf.gov/pubs/2020/nsf20108/nsf20108.jsp.
31 *Simplifying Access to Cloud Resources for Researchers: CloudBank*, Amazon Web Serv. (Nov. 16, 2020), https://aws.amazon.com/blogs/publicsector/simplifying-access-cloud-resources-researchers-cloudbank/.
32 *Community & Education Resource Requests*, *supra* note 11.
33 Larry Dignan, *AWS Cloud Computing Ops, Data Centers, 1.3 Million Servers Creating Efficiency Flywheel*, ZDNet (June 17, 2016), https://www.zdnet.com/article/aws-cloud-computing-ops-data-centers-1-3-million-servers-creating-efficiency-flywheel/; Rich Miller, *Ballmer: Microsoft Has 1 Million Servers*, Data Ctr. Knowledge (July 15, 2013), https://www.datacenterknowledge.com/archives/2013/07/15/ballmer-microsoft-has-1-million-servers; Daniel Oberhaus, *Amazon, Google, Microsoft: Here's Who Has the Greenest Cloud*, Wired (Dec. 18, 2019), https://www.wired.com/story/amazon-google-microsoft-green-clouds-and-hyperscale-data-centers/; Russell Brandom, *Mapping out Amazon's Invisible Server Empire*, The Verge (May 10, 2019), https://www.theverge.com/2019/5/10/18563485/amazon-web-services-internet-location-map-data-center.
34 *See, e.g., AWS Pricing*, Amazon Web Services, https://aws.amazon.com/pricing/; *Overview of Cloud Billing Concepts*, Google Cloud, https://cloud.google.com/billing/docs/concepts; *Azure Pricing*, Azure, https://azure.microsoft.com/en-us/pricing/#product-pricing.
35 Large research universities already negotiate enterprise agreements with cloud providers.
36 *What We Do*, XSEDE, https://www.xsede.org/about/what-we-do (last visited Sept. 19, 2021).
37 *XSEDE Overall Organization*, XSEDE Wiki, https://confluence.xsede.org/display/XT/XSEDE+Overall+Organization (last visited Sept. 19, 2021).
38 *XSEDE Allocations Info & Policies*, XSEDE, https://portal.xsede.org/allocations/policies (last visited Sept. 19, 2021).
39 *Id.*
40 *Id.*
41 *Startup Allocations*, XSEDE, https://portal.xsede.org/allocations/startup (last visited Sept. 19, 2021).
42 *Id.*
43 *Id.*
44 *Id.*
45 *Research Allocations*, XSEDE, https://portal.xsede.org/allocations/research (last visited Sept. 19, 2021).
46 *Id.*
47 *Id.*
48 *Id.*
49 *XSEDE Allocations Info & Policies*, *supra* note 36.
50 *XSEDE Campus Champions*, *supra* note 20.
51 *Id.*
52 *Id.*
53 *XSEDE as a Collaborator on Proposals*, XSEDE, https://www.xsede.org/about/collaborating-with-xsede (last visited Sept. 19, 2021).
54 *COVID-19 HPC Consortium*, XSEDE, https://www.xsede.org/covid19-hpc-consortium (last visited Sept. 19, 2021).
55 Amazon, for example, introduced its P4, P3, and P2 instances in 2020, 1997, and 1996, respectively. Frederic Lardinois, *AWS Launches Its Next-Gen GPU Instances with 8 Nvidia A100 Tensor Core GPUs*, TechCrunch (Nov. 2, 2020), https://social.techcrunch.com/2020/11/02/aws-launches-its-next-gen-gpu-instances/; Ian C. Schafer, *Amazon Elastic Compute Cloud P3 Launched alongside NVIDIA GPU Cloud*, SD Times (Oct. 26, 2017), https://sdtimes.com/ai/amazon-elastic-compute-cloud-p3-launched-alongside-nvidia-gpu-cloud/; Jeff Barr, *New P2 Instance Type for Amazon EC2 – Up to 16 GPUs*, Amazon Web Services (Sept. 29, 2016), https://aws.amazon.com/blogs/aws/new-p2-instance-type-for-amazon-ec2-up-to-16-gpus/. The introduction years of the P4 and P3 instances line up with the release of NVIDIA's newest general purpose data center GPUs.
56 *See, e.g.,* Sarah Wang & Martin Casado, *The Cost of Cloud, a Trillion Dollar Paradox*, Andreessen Horowitz (May 27, 2021), https://a16z.com/2021/05/27/cost-of-cloud-paradox-market-cap-cloud-lifecycle-scale-growth-repatriation-optimization/.
57 Preston Smith et al., *Community Clusters or the Cloud: Continuing Cost Assessment of On-Premises and Cloud HPC in Higher Education*, 2019 Proceedings Practice & Experience Advanced Res. Computing on Rise of the Machines 1 (2019). The amortized cost includes the annual compute cost, subsidized hardware cost, and power costs, but does not include personnel costs, as such costs are fixed and would be recurred regardless of whether a cluster existed physically on-prem or on the cloud. *Id.*
58 Craig A. Stewart et al., *Return on Investment for Three Cyberinfrastructure Facilities: A Local Campus Supercomputer; the NSF-Funded Jetstream Cloud System; and XSEDE*, 11 Int'l Conf. on Utility & Cloud Computing 223 (2018).
59 Srijith Rajamohan & Robert E. Settlage, *Informing the On/Off-prem Cloud Discussion in Higher Education*, 2020 Practice & Experience Adv. Res. Computing 64 (2020). The cost sources include hardware, software services, software administration, electricity, and facilities but do not include computational scientists support, scientific software licenses, and data transfer costs. The study is also limited to Virginia Tech's particular cloud workload.
60 Jennifer Villa & Dave Troiano, *Choosing Your Deep Learning Infrastructure: The Cloud vs. On-Prem Debate*, Determined AI (July 30, 2020), https://determined.ai/blog/cloud-v-onprem/; *Is HPC Going to Cost Me a Fortune?*, insideHPC, https://insidehpc.com/hpc-basic-training/is-hpc-going-to-cost-me-a-fortune/.
61 Interview with Suzanne Talon, Regional Director, Compute Canada (Jan. 14, 2021).
62 Compute Canada, Cloud Computing for Researchers 1 (2016), https://www.computecanada.ca/wp-content/uploads/2015/02/CloudStrategy2016-2019-forresearchersEXTERNAL-1.pdf.
63 *US Plans $1.8 Billion Spend on DOE Exascale Supercomputing*, HPCwire (Apr. 11, 2018), https://www.hpcwire.com/2018/04/11/us-plans-1-8-billion-spend-on-doe-exascale-supercomputing/; *Federal Government*, Advanced HPC, https://www.advancedhpc.com/pages/federal-government; *United States Continues To Lead World In Supercomputing*, Energy.gov, https://www.energy.gov/articles/united-states-continues-lead-world-supercomputing; *High Performance Computing*, Energy.gov, https://www.energy.gov/science/initiatives/high-performance-computing.
64 *See, e.g., DOE Announces Five New Energy Projects at LLNL*, LLNL (Nov. 13, 2020), https://www.llnl.gov/news/doe-announces-five-new-energy-projects-llnl; *New HPCMP System at the AFRL DSRC DoD Supercomputing Resource Center to Provide over Nine PetaFLOPS of Computing Power to*

Address Physics, AI, and ML Applications for DoD Users, DOD HPC, https://www.hpc.mil/images/hpcdocs/newsroom/21-19_TI-21_web_announcement_AFRL_DSRC.pdf; *Public Announcement*, DOD HPC, https://www.hpc.mil/images/hpcdocs/newsroom/awards_and_press/HC101321D0002_PUBLIC_ANNOUNCEMENT_20210505.pdf.

65 Devin Coldewey, *$600M Cray Supercomputer Will Tower Above the Rest — to Build Better Nukes*, TechCrunch (Aug. 13, 2019), https://social.techcrunch.com/2019/08/13/600m-cray-supercomputer-will-tower-above-the-rest-to-build-better-nukes/; *CORAL-2 RFP*, Oak Ridge Nat'l Laboratory (Apr. 9, 2018), https://procurement.ornl.gov/rfp/CORAL2/.

66 *See, e.g.*, *NSF Funds Five New XSEDE-Allocated Systems*, Nat'l Sci. Found. (Aug. 10, 2020), https://www.xsede.org/-/nsf-funds-five-new-xsede-allocated-systems.

67 Timothy Prickett Morgan, *Bending The Supercomputing Cost Curve Down*, The Next Platform (Dec. 2, 2019), http://www.nextplatform.com/2019/12/02/bending-the-supercomputing-cost-curve-down/; Ben Dickson, *The GPT-3 Economy*, TechTalks (Sept. 21, 2020), https://bdtechtalks.com/2020/09/21/gpt-3-economy-business-model/.

68 Elijah Wolfson, *The US Just Retook the Title of World's Fastest Supercomputer from China*, Quartz (June 9, 2018), https://qz.com/1301510/the-us-has-the-worlds-fastest-supercomputer-again-the-200-petaflop-summit/.

69 *November 2020*, TOP500 (Nov. 2020), https://www.top500.org/lists/top500/2020/11/.

70 *U.S. Department of Energy and Cray to Deliver Record-Setting Frontier Supercomputer at ORNL*, Oak Ridge Nat'l Laboratory (May 7, 2019), https://www.ornl.gov/news/us-department-energy-and-cray-deliver-record-setting-frontier-supercomputer-ornl.

71 Coury Turczyn, *Building an Exascale-Class Data Center*, Oak Ridge Leadership Computing Facility (Dec. 11, 2020), https://www.olcf.ornl.gov/2020/12/11/building-an-exascale-class-data-center/.

72 Don Clark, *Intel Slips, and a High-Profile Supercomputer Is Delayed*, N.Y. Times (Aug. 27, 2020), https://www.nytimes.com/2020/08/27/technology/intel-aurora-supercomputer.html; Mila Jasper, *10 of 15 of DOD's Major IT Projects Are Behind Schedule, GAO Found*, Nextgov (Jan. 4, 2021), https://www.nextgov.com/it-modernization/2021/01/10-15-dods-major-it-projects-are-behind-schedule-gao-found/171155/.

73 *See* Nattakarn Phaphoom et al., *A Survey Study on Major Technical Barriers Affecting the Decision to Adopt Cloud Services*, 103 J. Systems & Software 167, 171-72 (2015) (describing data portability, integration with existing systems, migration complexity, and availability as major barriers to cloud adoption); Abdulrahman Alharthi et al., *An Overview of Cloud Services Adoption Challenges in Higher Education Institutions*, 2 Proceedings of the Int'l Workshop on Emerging Software as a Service & Analytics 102, 107-08 (2015) (acknowledging the low rate of cloud computing adoption in higher education and emphasizing that bolstering both the perceived ease of use and the actual usefulness of cloud computing can increase the adoption rate).

74 *See* Dep't of Energy, FY 2021 Budget Justification Volume 4: Science (2020).

75 Joe Weinman, Cloudonomics: The Business Value of Cloud Computing (2012).

76 OLCF supports and manages ORNL's supercomputing resources, including Summit and eventually Frontier. This figure accounts for "operations and user support at the LCF facilities–including power, space, leases, and staff. *Id.* at 37-38.

77 ACLF supports and manages Argonne National Laboratory's computing resources, including the Theta system and, later this year, the new Aurora computer, another DOE exascale HPC system. *Id*.

78 OLCF operated its Titan HPC system for 7 years. *See* Coury Turczyn, *supra* note 72. ACLF also operated its Mira HPC system for 7 years. *Argonne's Mira Supercomputer to Retire After Years of Enabling Groundbreaking Science*, HPCwire (Dec. 20, 2019), https://www.hpcwire.com/2019/12/20/argonnes-mira-supercomputer-to-retire-after-years-of-enabling-groundbreaking-science/. If still operational, these systems would rank about the 19th and 29th fastest in the world, respectively. *Compare November 2020*, *supra* note 70, *with TOP500 List - June 2019*, TOP500 (June 2019), https://www.top500.org/lists/top500/list/2019/06/.

79 *See, e.g.*, Kim Zetter, *Top Federal Lab Hacked in Spear-Phishing Attack*, Wired (Apr. 20, 2011), https://www.wired.com/2011/04/oak-ridge-lab-hack/; Natasha Bertrand & Eric Wolff, *Nuclear Weapons Agency Breached amid Massive Cyber Onslaught*, Politico (Dec. 17, 2020), https://www.politico.com/news/2020/12/17/nuclear-agency-hacked-officials-inform-congress-447855 (last visited Mar. 2, 2021); Ryan Lucas, *List Of Federal Agencies Affected By A Major Cyberattack Continues To Grow*, NPR (Dec. 18, 2020), https://www.npr.org/2020/12/18/948133260/list-of-federal-agencies-affected-by-a-major-cyberattack-continues-to-grow (last visited Mar. 2, 2021).

80 We discuss data access models in Chapter Three.

81 *See Ongoing Projects*, RIKEN Ctr. for Computational Sci., https://www.r-ccs.riken.jp/en/fugaku/research/covid-19/projects/.

82 *Fugaku Retains Title as World's Fastest Supercomputer*, HPCWire (Nov. 17, 2020), https://www.hpcwire.com/off-the-wire/fugaku-retains-title-as-worlds-fastest-supercomputer/.

83 *November 2020*, *supra* note 70.

84 *Id.*

85 *Behind the Scenes of Fugaku as the World's Fastest Supercomputer*, Fujitsu (Feb. 2, 2021), https://blog.global.fujitsu.com/fgb/2021-02-02/behind-the-scenes-of-fugaku-as-the-worlds-fastest-supercomputer-1manufacturing/.

86 *Id.*

87 Don Clark, *Japanese Supercomputer Is Crowned World's Speediest*, N.Y. Times (June 22, 2020), https://www.nytimes.com/2020/06/22/technology/japanese-supercomputer-fugaku-tops-american-chinese-machines.html.

88 Justin McCurry, *Non-Woven Masks Better to Stop Covid-19, Says Japanese Supercomputer*, The Guardian (Aug. 26, 2020), http://www.theguardian.com/world/2020/aug/26/non-woven-masks-better-to-stop-covid-19-says-japanese-supercomputer.

89 *Fujitsu and RIKEN Complete Joint Development of Japan's Fugaku, the World's Fastest Supercomputer*, Fujitsu (Mar. 9, 2021), https://www.fujitsu.com/global/about/resources/news/press-releases/2021/0309-02.html.

90 *Id*.

91 *See, e.g.*, Rolf Harms & Michael Yamartino, The Economics of the Cloud (2010); Srijith Rajamohan & Robert E. Settlage, *Informing the On/Off-Prem Cloud Discussion in Higher Education*, 2020 Practice & Experience in Advanced Res. Computing 64 (2020); Byung Chul Tak et al., *To Move or Not To Move: The Economics of Cloud Computing*, 3 USENIX Conf. on Hot Topics in Cloud Computing 1 (2011); Edward Walker, Walter Brisken & Jonathan Romney, *To Lease or Not To Lease from Storage Clouds*, 43 Computer 44 (2010).

92 *See, e.g.*, Di Zhang et al., *RLScheduler: An Automated HPC Batch Job Scheduler Using Reinforcement Learning*, Cornell U. (Sept. 2, 2020), https://arxiv.org/pdf/1910.08925.pdf.

93 For instance, we have not been able to identify good estimates of electricity and cooling costs for DOE supercomputers.

94 Hugh Couchman et al., Compute Canada — Calcul Canada: A Proposal to the Canada Foundation for Innovation – National Platforms Fund

58 (2006).
95 *About*, Compute Canada, https://www.computecanada.ca/about/.
96 *National Systems*, Compute Canada, https://www.computecanada.ca/techrenewal/national-systems/.
97 *Compute Canada Technology Briefing,* Compute Canada (Nov. 2017), https://www.computecanada.ca/wp-content/uploads/2015/02/Technology-Briefing-November-2017.pdf.
98 *Cloud Computing for Researchers*, Compute Canada (Dec. 2016), https://www.computecanada.ca/wp-content/uploads/2015/02/CloudStrategy2016-2019-forresearchersEXTERNAL-1.pdf.
99 *Id.*
100 *Budget Submission 2018*, Compute Canada (2018), https://www.computecanada.ca/wp-content/uploads/2015/02/UTF-8Compute20Canada20Budget20Submission202018.pdf at 5.
101 Compute Canada projected it had only met about 55% of total demand for CPU compute hours in 2018. *Id.*
102 *Id.*
103 Compute Canada, Annual Report 2019-2020 4 (2020).
104 *Rapid Access Service*, Compute Canada, https://www.computecanada.ca/research-portal/accessing-resources/rapid-access-service/.
105 *Id.*
106 *Resource Allocation Competitions*, *supra* note 21.
107 *Id.*
108 *Id.*
109 *Id.*
110 *Id.*
111 *2021 Resource Allocations Competition Results*, *supra* note 21.

Chapter 3

1 *National Research Cloud Call to Action*, Stan. U. Inst. for Human-Centered Artificial Intelligence (2020), https://hai.stanford.edu/national-research-cloud-joint-letter.
2 We discuss the Privacy Act and privacy considerations in more detail in Chapter Five.
3 Amy O'Hara & Carla Medalia, *Data Sharing in the Federal Statistical System: Impediments and Possibilities,* 675 Annals Am. Acad. Pol. & Soc. Sci. 138, 140-41 (2018); *see also* President's Mgmt. Agenda, Federal Data Strategy 2020 Action Plan (2020).
4 Improved data access would, as we describe below, also promote evidence-based policymaking and improve trust in science (as data access makes replication efforts much easier).
5 *See, e.g.,* Nick Hart & Nancy Potok, Modernizing U.S. Data Infrastructure: Design Considerations for Implementing a National Secure Data Service to Improve Statistics and Evidence Building (2020).
6 These initiatives are successful in that they are sustainable and have been used by researchers to access multi-agency government data. The only exception is the National Secure Data Service (NSDS), which has not yet been implemented. We discuss the NSDS alongside the Census Bureau and the Evidence-Based Policy-Making Act of 2018 below. Importantly, our focus in these case studies is not to evaluate their efforts or measure their exact levels of success but to identify and understand some of the differences and similarities in the range of data-sharing efforts.
7 For instance, private sector data may facilitate research regarding social media use, internet behavior, or fill in gaps for federal statistics research through big data analysis. *See* Robert M. Groves & Brian A. Harris-Kojetin, *Using Private-Sector Data for Federal Statistics,* Nat'l Ctr. for Biotechnology Info. (Jan. 12, 2017), https://www.ncbi.nlm.nih.gov/books/NBK425876/.
8 *See, e.g., National Data Service,* Nat'l Data Serv., http://www.nationaldataservice.org; *The Open Science Data Cloud,* Open Sci. Data Cloud, https://www.opensciencedatacloud.org, *Harvard Dataverse*, Harv. Dataverse, https://dataverse.harvard.edu, *FigShare*, https://figshare.com.
9 Facebook Data for Research provides access to a variety of libraries, via in-house platforms. *See, e.g., Facebook Data For Good*, Facebook (2020), https://dataforgood.fb.com/; *What is the Facebook Ad Library and How do I Search it?*, Facebook (2021), https://www.facebook.com/help/259468828226154; *Facebook Disaster Maps Methodology*, Facebook (May 15, 2019), https://research.fb.com/facebook-disaster-maps-methodology/.
10 For example, Twitter has a Developer Portal that provides access to their API to allow researchers to use user data for noncommercial purposes. *See Twitter Developers*, Twitter (2021), https://developer.twitter.com/en/portal/petition/academic/is-it-right-for-you*; Take Your Research Further with Twitter Data*, Twitter (2021), https://developer.twitter.com/en/solutions/academic-research. Thus, uploading Twitter data to a separate Cloud may provide few incentives to researchers who can use the API route.
11 *See* Nat'l Acad. of Sci., Innovations in Federal Statistics 31-42 (2017).
12 *See* Jennifer M. Urban, Joe Karaganis & Brianna M. Schofield, Notice & Takedown in Everyday Practice 39 (2017) (illustrating the difficulty that online service providers face in manually evaluating a large volume of data for potential infringement; for example, one online service provider explained that "out of fear of failing to remove infringing material, and motivated by the threat of statutory damages, its staff will take "six passes to try to find the [identified content]."); *see also* Letter from Thom Tillis, Marsha Blackburn, Christopher A. Coons, Dianne Feinstein et. al, to Sundar Pichai, Chief Executive Officer, Google Inc. (Sept. 3, 2019), https://www.ipwatchdog.com/wp-content/uploads/2019/09/9.3-Content-ID-Ltr.pdf ("We have heard from copyright holders who have been denied access to Content ID tools, and as a result, are at a significant disadvantage to prevent repeated uploading of content that they have previously identified as infringing. They are left with the choice of spending hours each week seeking out and sending notices about the same copyrighted works, or allowing their intellectual property to be misappropriated.").
13 To illustrate the costs of implementing Content ID on a large-scale platform, Google announced in a report in 2016 that YouTube had invested more than $60 million in Content ID. *See* Google, How Google Fights Piracy 6 (2016).
14 *See, e.g., AWS Customer Agreement*, Amazon (Nov. 30, 2020), https://aws.amazon.com/agreement/.
15 For instance, across the 29 distinct agencies in the Department of Health and Human Services (HHS), data "are largely kept in silos with a lack of organizational awareness of what data are collected across the Department and how to request access. Each agency operates within its own statutory authority and each dataset can be governed by a particular set of regulations." U.S. Dep't of Health & Human Services, The State of Data Sharing at the U.S. Department of Health and Human Services 4 (2018).

16 *See, e.g., id.* at 8 ("HHS lacks consistent and standardized processes for one agency to request data from another agency.").
17 O'Hara & Medalia, *supra* note 3, at 140-41.
18 *See id.* at 142 ("Most [data-sharing] agreements rely heavily on interpersonal relationships and informal quid pro quo arrangements, handling data requests in a less centralized fashion.").
19 Jeffrey Mervis, *How Two Economists Got Direct Access to IRS Tax Records*, Sci. Mag. (May 22, 2014), https://www.sciencemag.org/news/2014/05/how-two-economists-got-direct-access-irs-tax-records.
20 *See* Robert M. Groves & Adam Neufeld, Accelerating the Sharing of Data Across Sectors to Advance the Common Good 17 (2017).
21 *See, e.g., Data Use Agreement*, Dep't Health & Human Services, https://www.hhs.gov/sites/default/files/ocio/eplc/EPLC%20Archive%20Documents/55-Data%20Use%20Agreement%20%28DUA%29/eplc_dua_practices_guide.pdf.
22 O'Hara & Medalia, *supra* note 3, at 138, 141.
23 Bipartisan Pol'y Ctr., Barriers to Using Government Data: Extended Analysis of the U.S. Commission on Evidence-Based Policymaking's Survey of Federal Agencies and Offices 18-20 (2018).
24 *See Research Data Assistance Center (ResDAC)*, Ctr. for Medicare & Medicaid Services (Aug. 30, 2018), https://www.cms.gov/Research-Statistics-Data-and-Systems/Research/ResearchGenInfo/ResearchDataAssistanceCenter.
25 O'Hara & Medalia, *supra* note 3, at 141.
26 Michelle Mello et al., *Waiting for Data: Barriers to Executing Data Use Agreements*, 367 Sci. Mag. 150 (Jan. 10, 2020), https://www.sciencemagazinedigital.org/sciencemagazine/10_january_2020/MobilePagedArticle.action?articleId=1552284#articleId1552284.
27 Interview with Amy O'Hara, Executive Director, Georgetown Federal Statistical Research Data Center (Apr. 22, 2021); *see also Special Sworn Research Program,* Bureau of Econ. Analysis, https://www.bea.gov/research/special-sworn-researcher-program; Nat'l Ctr. for Educ. Stat., Restricted-Use Data Procedures Manual (2011).
28 *See* U.S. Gov't Accountability Office, Federal Agencies Need to Address Aging Legacy Systems (2016).
29 O'Hara & Medalia, *supra* note 3, at 140-41.
30 *See, e.g., id.;* U.S. Gov't Accountability Office, *supra* note 28.
31 President's Mgmt. Agenda, *supra* note 3, at 11.
32 Groves & Neufeld, *supra* note 20, at 12-13. For a precise definition of sensitive data, see *Glossary: Sensitive Information*, Nat'l Inst. Standards & Tech., https://csrc.nist.gov/glossary/term/sensitive_information.
33 Shanna Nasiri, *FedRAMP Low, Moderate, High: Understanding Security Baseline Levels*, Reciprocity (Sept. 24, 2019), https://reciprocity.com/fedramp-low-moderate-high-understanding-security-baseline-levels/.
34 Michael McLaughlin, *Reforming FedRAMP: A Guide to Improving the Federal Procurement and Risk Management of Cloud Services*, Info. Tech. & Innovation Found. (June 15, 2020), https://itif.org/publications/2020/06/15/reforming-fedramp-guide-improving-federal-procurement-and-risk-management.
35 *Frequently Asked Questions*, FedRAMP, https://www.fedramp.gov/faqs.
36 *Do Once, Use Many - How Agencies Can Reuse a FedRAMP Authorization*, FedRAMP (May 7, 2020), https://www.fedramp.gov/how-agencies-can-reuse-a-fedramp-authorization/.
37 FedRAMP, FedRAMP Low, Moderate, and High Security Control Baselines (2021).
38 *Security and Privacy Controls for Information Systems and Organizations*, Nat'l Inst. Standards & Tech. (Sept. 23, 2020), https://csrc.nist.gov/publications/detail/sp/800-53/rev-5/final.
39 *See, e.g., id.*; *NIST Risk Management Framework AC-2: Account Management*, Nat'l Inst. Standards & Tech., https://csrc.nist.gov/Projects/risk-management/sp800-53-controls/release-search#!/control?version=4.0&number=AC-2; *NIST Risk Management Framework AC-3: Access Enforcement*, Nat'l Inst. Standards & Tech., https://csrc.nist.gov/Projects/risk-management/sp800-53-controls/release-search#!/control?version=5.1&number=AC-3.
40 *See* Mark Bergen, *Google Engineers Refused to Build Security Tool to Win Military Contracts*, Bloomberg (June 21, 2018), https://www.bloomberg.com/news/articles/2018-06-21/google-engineers-refused-to-build-security-tool-to-win-military-contracts.
41 *See* Nat'l Inst. Standards & Tech., Standards for Security Categorization of Federal Information and Information Systems (2004).
42 *Partnering with FedRAMP*, FedRAMP, https://www.fedramp.gov/cloud-service-providers/. While it may cost cloud service providers between $365,000 and $865,000 and take 6-12 months to receive FedRAMP compliance, Adam Isles, Securing Your Cloud Solutions: Research and Analysis on Meeting FedRAMP/Government Standards 21 (2017), such costs are borne by the cloud service providers themselves, not the providers' customers. Indeed, FedRAMP uses a "do once, use many" model: Once a cloud service provider obtains an authorization to operate (ATO), that ATO can be leveraged and reciprocated across multiple customers, eliminating duplicative efforts and inconsistencies that would come from requiring multiple re-authorizations. *Id.* at 11.
43 Even within FedRAMP there are substantial amounts of variation in how different organizations ensure compliance with the relevant controls and standards, with many of the controls written broadly enough to give room for substantial interpretation. However, it does lay out a variety of considerations and requirements that are consistent across domains and allows a degree of predictability and reliance that is not present in other aspects of federal data governance.
44 O'Hara & Medalia, *supra* note 3, at 141 ("Data sharing is taking place on a mandatory or voluntary basis, and data requests are managed through a designated staff/process or diffusely through an organization.").
45 Bipartisan Pol'y Ctr., *supra* note 23, at 17 ("The lack of standard procedures or guidelines for sharing data across federal agencies that fund research makes efforts to link and share data difficult or inefficient.").
46 *See, e.g.,* Amy O'Hara*, US Federal Data Policy: An Update on The Federal Data Strategy and The Evidence Act,* 5 Int'l J. Population Data Sci. 5 (2020).
47 While existing federal efforts and initiatives are already aimed at harmonizing data sharing best practices, *see, e.g.*, *2020 Action Plan,* Federal Data Strategy (May 14, 2020), https://strategy.data.gov/action-plan/, the NRC can accelerate these efforts. Indeed, the development of clear, consistent standards is crucial in facilitating data-sharing. David Crotty, Ida Sim & Michael Stebbins, Open Access to Federally Funded Research Data 7 (2020).
48 These requirements are inconsistent and out-of-date due to difficulties in defining risk as well as risk aversion on the parts of agencies. *See* O'Hara & Medalia, *supra* note 3, at 140-41; *see also* David S. Johnson et al., *The Opportunities and Challenges of Using Administrative Data Linkages to Evaluate Mobility*, 657 Annals Am. Acad. Pol. & Soc. Sci. 252-53 (2015).
49 For a discussion of inference threats, see Nat'l Acad. of Sci., Eng'g & Med., Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps 68 (2017).
50 Congzheng Song & Ananth Raghunathan, *Information Leakage in Embedding Models*, Cornell U. (Mar. 31, 2020), https://arxiv.org/abs/2004.00053.
51 *See, e.g.*, *Statistical Safeguards*, Census Bureau (July 1, 2021), https://www.census.gov/about/policies/privacy/statistical_safeguards.html.

52 Alexandra Wood et al., *Differential Privacy: A Primer for a Non-Technical Audience*, 21 Vand. J. Ent. & Tech. L. 209 (2018).
53 *Regulating Access to Data*, UK Data Serv., https://www.ukdataservice.ac.uk/manage-data/legal-ethical/access-control/five-safes.
54 *Administrative Data Research Facility*, Coleridge Initiative, https://coleridgeinitiative.org/adrf/.
55 *See* O'Hara, *supra* note 46.
56 For additional discussion of the privacy implications of the NRC, see Chapter Five.
57 *See* U.S. Office of Mgmt. & Budget, Barriers to Using Administrative Data for Evidence-Building 7 (2016).
58 *Administrative Data Research Facility*, *supra* note 54.
59 *Id.*
60 *Training*, Coleridge Initiative, https://coleridgeinitiative.org/training/.
61 *ADRF User Guide: Data Explorer*, Coleridge Initiative, https://coleridgeinitiative.org/adrf/documentation/using-the-adrf/data-explorer/.
62 *ADRF User Guide: Exporting Results*, Coleridge Initiative, https://coleridgeinitiative.org/adrf/documentation/using-the-adrf/exporting-results/.
63 *Id.*
64 *ADRF User Guide: Data Hashing Application*, Coleridge Initiative, https://coleridgeinitiative.org/adrf/documentation/adrf-overview/data-hashing-application/.
65 *ADRF User Guide: Security Model and Compliance*, Coleridge Initiative, https://coleridgeinitiative.org/adrf/documentation/adrf-overview/security-model-and-compliance/.
66 *Overview for Collaborators*, Coleridge Initiative, https://coleridgeinitiative.org/collaborators/.
67 *Data*, Stan. Med. Ctr. for Population Health Sci., https://med.stanford.edu/phs/data.html.
68 Stanford Ctr. for Philanthropy & Civ. Soc'y, Trusted Data Intermediaries 2-3 (2018).
69 Others have also recognized the benefit of universal DUA templates. *See* Mello et al., *supra* note 26, at 150; *Guidance for Providing and Using Administrative Data for Statistical Purposes*, Office of Mgmt. & Budget (Feb. 14, 2014), https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2014/m-14-06.pdf.
70 *Data | Center for Population Health Sciences | Stanford Medicine*, Stan. Med. Ctr. for Population Health Sci., https://med.stanford.edu/phs/data.html.
71 *See Stanford PHS – Datasets*, Redivis, https://redivis.com/StanfordPHS/datasets?orgDatasets-tags=109.medicare.
72 *Access Levels*, Redivis (July 2020), https://docs.redivis.com/reference/data-access/access-levels.
73 *Step 1: Getting Access*, Stan. Med. PHS Documentation, https://phsdocs.developerhub.io/start-here/getting-data-access.
74 *Id.*
75 *Id.*
76 *PHS Data-Use Workflow*, Stan. Med. PHS Documentation, https://phsdocs.stanford.edu/start-here/phs-data-use-workflow.
77 *Id.*
78 *PHS Computing Environment*, Stan. Med. PHS Documentation, https://phsdocs.stanford.edu/computing-environment.
79 *Id.*
80 *See* U.S. Gov't Accountability Office, Federal Agencies Need to Address Aging Legacy Systems 15 (2016) (noting that from 2010-2015, many federal agencies increased their spending on operations and maintenance due to legacy systems).
81 David Freeman Engstrom, Daniel E. Ho, Catherine M. Sharkey & Mariano-Florentino Cuéllar, Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies 6, 71-72 (2020).
82 *Id.* at 6-7.
83 *Id.* at 71-72.
84 *Id.* at 73.
85 *Id.* at 6.
86 *See* Results for America, The Promise of the Foundations for Evidence-Based Policymaking Act and Proposed Next Steps (2019)
87 For example, the Uniform Federal Crime Reporting Act of 1988 requires federal law enforcement agencies to share crime data with the FBI. *See* 34 U.S.C. §§41303(c)(2), (3), (4). Unfortunately, though, no federal agencies apparently currently share their data with the FBI under this law. Nat'l Acad. of Sci., *supra* note 11, at 41 (2017).
88 Katharine G. Abraham & Ron Haskins, The Promise of Evidence-Based Policymaking; Comm'n on Evidence-Based Policymaking (2018).
89 These privacy-preserving mechanisms are especially important in light of ongoing legal and political challenges in differential privacy application to Federal data. *See, e.g.,* Dan Bouk & danah boyd, Democracy's Data Infrastructure (2021).
90 Foundations for Evidence-Based Policymaking Act of 2018, Pub. L. No. 115-435.
91 Confidential Information Protection and Statistical Efficiency Act of 2002, Pub. L. No. 107-347.
92 *Overview*, Federal Data Strategy (2020), https://strategy.data.gov/overview/.
93 *UK Data Service*, UK Data Serv., https://www.ukdataservice.ac.uk/ (last visited Jun. 21, 2021).
94 For example, the Social Security Administration alone has over 14 petabytes of data, stored in roughly 200 databases. Engstrom, Ho, Sharkey & Cuéllar, *supra* note 81, at 72.
95 *Google Earth Engine*, Google Earth Engine, https://earthengine.google.com (last visited Aug. 15, 2021).
96 *World of Work*, ADR UK, https://www.adruk.org/our-work/world-of-work/.
97 *Annual Respondents Database, 1973-2008: Secure Access*, UK Data Serv. (2020), https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=6644.
98 *UK Innovation Survey*, UK Data Serv. (2021), https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=6699.
99 *Quarterly Labour Force Survey, 1992-2021: Secure Access*, UK Data Serv. (2021), https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=6727.
100 *Understanding Society: Waves 1-10, 2009-2019 and Harmonised BHPS: Waves 1-18, 1991-2009: Secure Access*, UK Data Serv. (2021), https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=6676.
101 These datasets have helped researchers tackle some specific, public good questions. *See, e.g.,* Francisco Perales, *Why Does the Work Women Do Pay Less Than the Work Men Do?*, UK Data Serv. (Dec. 8, 2011), https://beta.ukdataservice.ac.uk/impact/case-studies/case-study?id=62; Eva-Maria Bonin, *Do Parenting Programmes Reduce Conduct Disorder?*, UK Data Serv. (Apr. 4, 2012), https://beta.ukdataservice.ac.uk/impact/case-studies/case-study?id=93.
102 *Identifying Priority Access or Quality Improvements for Federal Data and Models for Artificial Intelligence Research and Development (R&D), and Testing; Request for Information*, 84 Fed. Reg. 32962 (July 10, 2019).

103 Nick Hart, *Data Coalition Comments on AI Data and Model R&D RFI*, Data Coalition (Aug. 9, 2019), http://www.datacoalition.org/wp-content/uploads/2019/09/Comment.RFI_.OMB_.2019-14618.DataCoalition.pdf.
104 *Id.*
105 *See* Adam R. Pah et al., *How to Build a More Open Justice System*, 369 Sci. 134 (2020); *see also* Seamus Hughes, *The Federal Courts Are Running an Online Scam*, Politico (Mar. 20, 2019), https://www.politico.com/magazine/story/2019/03/20/pacer-court-records-225821/.
106 *Legal Authority and Policies for Data Linkage at Census*, Census Bureau (Apr. 4, 2018), https://www.census.gov/about/adrm/linkage/about/authority.html.
107 *BLS Restricted Data Access*, U.S. Bureau of Lab. Stat., https://www.bls.gov/rda/restricted-data.htm (last updated May 20, 2021).
108 *Welcome to the PDS*, NASA, https://pds.nasa.gov.

Chapter 4

1 While we believe that these are the primary axes for consideration, some secondary considerations include organizational clout, talent retention, and bureaucratic overhead.
2 Congressional Research Serv., Federally Funded Research and Development Centers (FFRDCs): Background and Issues for Congress 1 (2020).
3 *Id*. *See also About IDA*, Inst. Defense Analyses, https://www.ida.org/about-ida (emphasizing that IDA, the private sector subcontractor that operates the Science & Technology Policy Institute and several other FFRDCs, "enjoys unusual access to classified government information and sensitive corporate proprietary information."); U.S. Gov't Accountability Office, Federally Funded Research and Development Centers: Improved Oversight and Evaluation Needed for DOD's Data Access Pilot Program 6 (2020) (discussing how the Department of Defense was able to establish a three-year pilot program that allowed its FFRDC researchers to forgo having to obtain nondisclosure agreements with each data owner in order to streamline the data-access process).
4 Nick Hart & Nancy Potok, Modernizing U.S. Data Infrastructure: Design Considerations for Implementing a National Secure Data Service to Improve Statistics and Evidence Building (2020).
5 *Id.* at 26.
6 *Id.* at 26-27, 29-30.
7 U.S. Gov't Accountability Office, *supra* note 3, at 6. Note that while the FFRDC must operate to serve its sponsors, in establishing an FFRDC, the sponsor must ensure that it operates with substantial independence; the FFRDC must be "operated, managed, or administered by an autonomous organization or as an identifiably separate operating unit of a parent organization." *See* Federal Acquisition Regulations [hereinafter "FAR"] § 35.017(a)(2).
8 One example of this is the Science & Technology Policy Institute, which we discuss in a case study below.
9 U.S. Dep't of Energy, The State of the DOE National Laboratories 11-13 (2020).
10 *See, e.g., More Federal Agencies Head to the Cloud With Azure Government*, Applied Info. Sci. (Feb. 23, 2018), https://www.ais.com/more-federal-agencies-head-to-the-cloud-with-azure-government/; *see also AWS GovCloud*, Amazon, https://aws.amazon.com/govcloud-us/. Microsoft was also previously awarded a $10 billion contract from the Pentagon. *See* Kate Conger, *Microsoft Wins Pentagon's $10 Billion JEDI Contract, Thwarting Amazon*, N.Y. Times (Sept. 4, 2020), https://www.nytimes.com/2019/10/25/technology/dod-jedi-contract.html. However, this contract was recently canceled "due to evolving requirements, increased cloud conservancy and industry advances." Ellie Kaufman & Zachary Cohen, *Pentagon Cancels $10 Billion Cloud Contract Given to Microsoft Over Amazon*, CNN (July 6, 2021), https://www.cnn.com/2021/07/06/tech/defense-department-cancels-jedi-contract-amazon-microsoft/index.html. The Pentagon will now instead seek new bids for an updated Joint Warfighting Cloud Capability (JWCC) contract from Amazon and Microsoft. *Id.*
11 *See, e.g.,* Bram Bout, *Helping Universities Build What's Next with Google Cloud Platform*, Google (Oct. 25, 2016), https://blog.google/outreach-initiatives/education/helping-universities-build-whats-next-google-cloud-platform; *Cloud Computing for Education*, Amazon, https://aws.amazon.com/education/.
12 Congressional Research Serv., *supra* note 2, at 11-12 (2020).
13 U.S. Dep't of Energy, Annual Report on the State of the DOE National Laboratories 87 (2017).
14 Congressional Research Serv., *supra* note 2, at 19.
15 Congressional Research Serv., Office of Science and Technology Policy (OSTP): History and Overview 9 (2020). STPI's duties are also specified in 42 U.S.C. § 6686.
16 *What are FFRDCs?*, Inst. Defense Analyses, https://www.ida.org/ida-ffrdcs.
17 *Id.*
18 *Sponsors*, Inst. Defense Analyses, https://www.ida.org/en/about-ida/sponsors.
19 *Id.*
20 Congressional Research Serv., *supra* note 15, at 9-10.
21 For instance, from 2008-2012, these other federal agencies contributed a total of $9.8 million of funding to STPI while NSF contributed about $24 million. U.S. Gov't Accountability Office, Federally Funded Research Centers: Agency Reviews of Employee Compensation and Center Performance 43-44 (2014).
22 Congressional Research Serv., *supra* note 15, at 9-10.
23 *Id.*
24 42 U.S.C. § 6686(d).
25 42 U.S.C. § 6686(e).
26 Sci. & Tech. Pol'y Inst., Report to the President Fiscal Year 2020 (2020).
27 *See, e.g., Open Government*, Millennium Challenge Corp., https://www.mcc.gov/initiatives/initiative/open; Nat'l Geospatial Advisory Comm., Advancing the National Spatial Data Infrastructure Through Public-Private Partnerships and Other Innovative Partnerships (2020); Nat'l Aeronautics & Space Admin., Public-Private Partnerships for Space Capability Development 33-36 (2014).
28 *Big Data Value Public-Private Partnership*, European Comm'n (Mar. 9, 2021), https://digital-strategy.ec.europa.eu/en/library/big-data-value-public-private-partnership.
29 RAND, Public-Private Partnerships for Data-Sharing: A Dynamic Environment 33, 99 (2000).
30 *See Homepage - Alberta Data Partnerships*, Alberta Data Partnerships, http://abdatapartnerships.ca (last visited Aug. 15, 2021).

31 Alberta Data Partnerships, A P3 Success Story 1 (2017).
32 *Id.*
33 *Id.* at 19, 35.
34 *Id.* at 15.
35 *Id.*
36 *Id.* at 1.
37 Nat'l Geospatial Advisory Comm., Public-Private Partnership Use Case: Alberta Data Partnerships 1 (2020).
38 Alberta Data Partnerships, *supra* note 31, at 15.
39 *Id.* at 16.
40 *The COVID-19 High Performance Computing Consortium*, COVID-19 HPC Consortium, https://covid19-hpc-consortium.org.
41 *Id.*
42 *See, e.g.,* David Hall, Why Public-Private Partnerships Don't Work (2015); *Disadvantages and Pitfalls of the PPP Option*, APMG Int'l, https://ppp-certification.com/ppp-certification-guide/54-disadvantages-and-pitfalls-ppp-option.
43 Graeme A. Hodge, Carsten Greve & Anthony E. Boardman, International Handbook on Public-Private Partnerships, 187-90 (2012).
44 For example, on one end of a spectrum, the California Teale Data Center creates, owns, maintains, and archives its own datasets for private sector use. In contrast, the Pennsylvania Spatial Data Access houses metadata, requiring users to ask the actual data sources for access. RAND, *supra* note 29, at 102-03. We encourage the Task Force to examine this comprehensive report to assess the various organizational options for a PPP data clearinghouse model.
45 Angela Ballantyne & Cameron Stewart, *Big Data and Public-Private Partnerships in Healthcare and Research*, 11 Asian Bioethics R. 315, 315 (2019).
46 *See* Gov't Accountability Office, Human Capital: Improving Federal Recruiting and Hiring Efforts; *see also Catch and Retain: Improving Recruiting and Retention at Government Agencies*, Salesforce, https://www.salesforce.com/solutions/industries/government/resources/government-recruitment-software/.
47 Partnership for Public Service, Survey on the Future of Government Service 2 (2020).
48 *Id.*

Chapter 5

1 *National Research Cloud Call to Action*, Stan. U. Inst. for Human-Centered Artificial Intelligence (2020), https://hai.stanford.edu/national-research-cloud-joint-letter.
2 Sensitive information, as defined by the National Institute of Standards and Technology, is information where the loss, misuse, or unauthorized access or modification could adversely affect the national interest or the conduct of federal programs, or the privacy to which individuals are entitled under 5 U.S.C. § 552a (the Privacy Act); that has not been specifically authorized under criteria established by an Executive Order or an Act of Congress to be kept classified in the interest of national defense or foreign policy. *See Glossary: Sensitive Information*, Nat'l Inst. Standards & Tech., https://csrc.nist.gov/glossary/term/sensitive_information.
3 We thank Mark Krass for these insights.
4 Agencies covered by the Act include "any Executive department, military department, Government corporation, Government controlled corporation, or other establishment in the executive branch of the [federal] Government (including the Executive Office of the President), or any independent regulatory agency." 5 U.S.C. § 552(f)(1).
5 U.S. General Accounting Office, Record Linkage and Privacy: Issues in Creating New Federal Research and Statistical Information 10 (2001).
6 Interview with Marc Groman, Former Senior Advisor for Privacy, White House Office of Management and Budget (Feb. 18, 2021); *see also* Bipartisan Pol'y Ctr., Barriers to Using Government Data: Extended Analysis of the U.S. Commission on Evidence-Based Policymaking's Survey of Federal Agencies and Offices 10 (2018).
7 *See* Joseph Near & David Darais, *Differentially Private Synthetic Data*, Nat'l Inst. Standards & Tech. (May 3, 2021), https://www.nist.gov/blogs/cybersecurity-insights/differentially-private-synthetic-data; *see also* Steven M. Bellovin et al., *Privacy and Synthetic Datasets*, 22 Stan. L. Rev. 1 (2019).
8 E-Government Act of 2002, Pub. L. No. 107-347.
9 Confidential Information Protection and Statistical Efficiency Act of 2002, 44 U.S.C. § 3501 (2012).
10 Foundations for Evidence-Based Policymaking Act of 2017, Pub. L. No. 115-435, 132 Stat. 5529 (2019).
11 President's Mgmt. Agenda, Federal Data Strategy 2020 Action Plan (2020).
12 Privacy Act of 1974, 5 U.S.C. § 552a (2012).
13 There are many versions of the Fair Information Practice Principles, and the U.S. government has not institutionalized a specific version, though the version used by the Department of Homeland Security is commonly referenced (available at: https://www.dhs.gov/publication/privacy-policy-guidance-memorandum-2008-01-fair-information-practice-principles). The Organisation for Economic Cooperation and Development produced an influential version of them in 1980 (revised in 2013), which remains an authoritative source. *OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*, OECD (2013), https://www.oecd.org/digital/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm.
14 *See* David Freeman Engstrom, Daniel E. Ho, Catherine M. Sharkey & Mariano-Florentino Cuéllar, Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies (2020) (documenting present use of AI by government agencies).
15 5 U.S.C. § 552a (a)(5).
16 5 U.S.C. §§ 552a(a)(8)(A)(i)(I), (II).
17 *The Privacy Act of 1974*, Elec. Privacy Info. Ctr., https://epic.org/privacy/1974act/ (last visited Aug. 15, 2021).
18 *See Fact Sheet: National Secure Data Service Act Advances Responsible Data Sharing in Government,* Data Coalition (May 13, 2021), https://www.datacoalition.org/fact-sheet-national-secure-data-service-act-advances-responsible-data-sharing-in-government/; U.S Gov't Accountability Office, Record Linkage and Privacy: Issues in Creating New Federal Research and Statistical Information (2001).
19 It is no small irony that private companies in the U.S. have fulfilled that mission today. In fact, the U.S. government now approaches private industry, either through legal process or through procurement, when it requires data about individuals that the government itself does not collect. Senator Ron

Wyden has proposed legislation to prevent the government from making these purchases. *Wyden, Paul and Bipartisan Members of Congress Introduce The Fourth Amendment Is Not For Sale Act*, Ron Wyden U.S. Senator for Or. (Apr. 21, 2021), https://www.wyden.senate.gov/news/press-releases/wyden-paul-and-bipartisan-members-of-congress-introduce-the-fourth-amendment-is-not-for-sale-act-.

20 *See, e.g.*, World Econ. Forum, The Next Generation of Data-Sharing in Financial Services (2019).

21 *See, e.g.*, Stacie Dusetzina et al., *Linking Data for Health Services Research: A Framework and Instructional Guide,* Agency for Healthcare Research & Quality (Sept. 1, 2014), https://www.ncbi.nlm.nih.gov/books/NBK253315/.

22 *See, e.g.,* European Comm'n, A European Strategy for Data (2020) (arguing for cross-border data aggregation and linkage of both private and public sector data); M Sanni Ali et al., *Administrative Data Linkage in Brazil: Potentials for Health Technology Assessment*, 10 Frontiers in Pharmacology 984 (2019); *Data Linkage*, Australian Inst. of Health & Welfare (Jan. 4, 2020), https://www.aihw.gov.au/our-services/data-linkage.

23 *See, e.g.*, Elsa Augustine, Vikash Reddy & Jesse Rothstein, Linking Administrative Data: Strategies and Methods (2018) (describing tips for conducting data linkages in California); *see also* U.S. Dep't of Health & Human Services, Status of State Efforts to Integrate Health and Human Services Systems and Data (2016).

24 Ben Moscovitch, *How President Biden Can Improve Health Data Sharing For COVID-19 And Beyond*, Health Affairs (Mar. 1, 2021), https://www.healthaffairs.org/do/10.1377/hblog20210223.611803/full/.

25 *Home*, Johns Hopkins Coronavirus Resource Ctr., https://coronavirus.jhu.edu/.

26 The COVID Tracking Project, https://covidtracking.com/.

27 Fred Bazzoli, *COVID-19 Emergency Shows Limitations of Nationwide Data Sharing Infrastructure*, Healthcare IT News (June 2, 2020), https://www.healthcareitnews.com/news/covid-19-emergency-shows-limitations-nationwide-data-sharing-infrastructure.

28 *See, e.g.,* C. Jason Wang et al., *Response to COVID-19 in Taiwan: Big Data Analytics, New Technology, and Proactive Testing*, JAMA (Mar. 3, 2020), https://jamanetwork.com/journals/jama/fullarticle/2762689/; Fang-Ming Chen et al., *Big Data Integration and Analytics to Prevent a Potential Hospital Outbreak of COVID-19 in Taiwan*, 54 J. Microbiology, Immunology & Infection 129-30 (2020).

29 *See, e.g., Q&A on the Pentagon's "Total Information Awareness" Program*, Am. C.L. Union, https://www.aclu.org/other/qa-pentagons-total-information-awareness-program; *The Five Problems with CAPPS II: Why the Airline Passenger Profiling Proposal Should Be Abandoned*, Am. C.L. Union, https://www.aclu.org/other/five-problems-capps-ii.

30 *See, e.g.,* Barton Gellman, Dark Mirror: Edward Snowden and the American Surveillance State (2020); Edward Snowden, Permanent Record (2019).

31 5 U.S.C. § 552(b).

32 5 U.S.C. § 552(b)(3).

33 The Privacy Act also contains specific carve-outs for disclosures to the Census Bureau and to the National Archives and Records Administration. However, the carve-outs for these two agencies require that the disclosures be made for the purposes of a census survey and of recording historical value, respectively. Because the NRC's explicit purpose is to democratize AI innovation, it is unlikely that the NRC can take advantage of this existing exception to dataset disclosures under the Privacy Act.

34 For example, the Federal Emergency Management Agency's list of routine uses includes broad disclosure "[t]o an agency or organization for the purpose of performing audit or oversight operations as authorized by law, but only such information as is necessary and relevant to such audit or oversight function." Privacy Act of 1974; Department of Homeland Security Federal Emergency Management Agency-008 Disaster Recovery Assistance Files System of Records, 78 Fed. Reg. 25282 (May 30, 2013).

35 *See, e.g., Britt v. Naval Investigative Service*, 886 F.2d 544 (3d Cir. 1989).

36 *The Privacy Act of 1974*, *supra* note 17.

37 5 U.S.C. § 552(b)(5).

38 5 U.S.C. §§ 552a(a)(8)(B)(i), (ii) (emphasis added).

39 44 U.S.C. § 3561(8), (12).

40 U.S. Dep't of Health & Human Services, The State of Data Sharing at the U.S. Department of Health and Human Services 16 (2018).

41 44 U.S.C. § 3575(4).

42 *See* Engstrom, Ho, Sharkey & Cuéllar, *supra* note 14, at 16 (finding that the Bureau of Labor Statistics is one of the top ten agencies that use artificial intelligence); *Machine Learning*, Census Bureau (Apr. 17, 2019), https://www.census.gov/topics/research/data-science/about-machine-learning.html (asserting that the Census Bureau "needs" machine learning capabilities); Bureau of Econ. Analysis, 2020 Strategic Action Plan 7 (2020) (highlighting the importance of artificial intelligence and machine learning to BEA's strategy).

43 Group level data analyses also have inherent privacy risks and harms. *See, e.g.*, Linnet Taylor, *Safety in Numbers? Group Privacy and Big Data Analytics in the Developing World*, *in* Group Privacy: New Challenges of Data Technologies 13 (2017).

44 *See* 34 U.S.C. §§ 41303(c)(2), (3), (4).

45 Nat'l Acad. of Sci., Innovations in Federal Statistics 41 (2017).

46 *See* 13 U.S.C. § 6.

47 Nat'l Acad. of Sci., *supra* note 45, at 40.

48 According to the study, "an agency's legal counsel may advise against sharing data as a precautionary measure rather than because of an explicit prohibition." U.S. Gov't Accountability Office, Sustained and Coordinated Efforts Could Facilitate Data Sharing While Protecting Privacy 1 (2013).

49 *See* Amy O'Hara & Carla Medalia, *Data Sharing in the Federal Statistical System: Impediments and Possibilities*, 675 Annals Am. Acad. Pol. & Soc. Sci. 138, 141 (2018).

50 Robert M. Groves & Adam Neufeld, Accelerating the Sharing of Data Across Sectors to Advance the Common Good 12 (2017).

51 Bipartisan Pol'y Ctr., *supra* note 6, at 18-20.

52 *See* O'Hara & Medalia, *supra* note 49, at 141.

53 *Administrative Data Research UK*, ADR UK, https://www.adruk.org.

54 *About ADR UK*, ADR UK, https://www.adruk.org/about-us/about-adr-uk/.

55 *Id.*

56 *See World of Work,* ADR UK, https://www.adruk.org/our-work/world-of-work/.

57 *Funding Opportunities*, ADR UK, https://www.adruk.org/news-publications/funding-opportunities/.

58 *Id.*

59 *Id.*
60 *Funding Opportunity: A Unique Chance to Shape Data Science at the Heart of UK Government*, ADR UK (Apr. 8, 2021), https://www.adruk.org/news-publications/news-blogs/funding-opportunity-a-unique-chance-to-shape-data-science-at-the-heart-of-uk-government-384/.
61 *Funding Opportunities*, *supra* note 57.
62 Digital Economy Act 2017 (Gr. Br.).
63 ADR UK, Trust, Security and Public Interest: Striking the Balance 28 (2020).
64 *Id.*
65 *Id.*
66 *How Do We Work with Researchers?*, ADR UK, https://www.adruk.org/our-mission/working-with-researchers/.
67 *Accessing Secure Research Data as an Accredited Researcher*, Off. for Nat'l Stat., https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/approvedresearcherscheme.
68 *See* Nick Hart & Nancy Potok, Modernizing U.S. Data Infrastructure: Design Considerations for Implementing a National Secure Data Service to Improve Statistics and Evidence Building 17, 21 (2020).
69 *Id.*
70 *Id.* at 15.
71 President's Mgmt. Agenda, *supra* note 11, at 9.
72 *Id.* at 31.
73 *See What is Open Data?*, Open Data Handbook, https://opendatahandbook.org/guide/en/what-is-open-data/.

Chapter 6

1 *See Keeping Secrets: Anonymous Data Isn't Always Anonymous*, Berkeley Sch. of Info. (Mar. 15, 2014), https://ischoolonline.berkeley.edu/blog/anonymous-data/; Arvind Narayanan & Vitaly Shmatikov, *How to Break Anonymity of the Netflix Prize Dataset*, Cornell U. (Nov. 22, 2007), https://arxiv.org/pdf/cs/0610105.pdf.
2 Matt Fredrikson, Somesh Jha & Thomas Ristenpart, *Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures,* 22 Proceedings of the ACM Special Interest Group on Security, Audit & Control 1322 (2015); Nicholas Carlini et al., *Extracting Training Data from Large Language Models*, Cornell U. (June 15, 2021), https://arxiv.org/pdf/2012.07805.pdf.
3 *See, e.g., HIPAA Training, Certification, and Compliance*, HIPAA Training, https://www.hipaatraining.com/; *Research Data Management*, UK Data Serv., https://ukdataservice.ac.uk/learning-hub/research-data-management/.
4 Ashwin Machanavajjhala et al., *L-Diversity: Privacy Beyond K-Anonymity*, 22 Int'l Conf. Data Eng'g 24 (2006).
5 Cynthia Dwork & Aaron Roth, The Algorithmic Foundations of Differential Privacy (2014).
6 *See, e.g.,* Tara Bahrampour & Marissa J. Lang, *New System to Protect Census Data May Compromise Accuracy, Some Experts Way*, Wash. Post (June 1, 2021), https://www.washingtonpost.com/local/social-issues/2020-census-differential-privacy-ipums/2021/06/01/6c-94b46e-c30d-11eb-93f5-ee9558eecf4b_story.html; Kelly Percival, *Court Rejects Alabama Challenge to Census Plans for Redistricting and Privacy*, Brennan Ctr. (June 30, 2021), https://www.brennancenter.org/our-work/analysis-opinion/court-rejects-alabama-challenge-census-plans-redistricting-and-privacy.
7 *See, e.g.,* Leonard E. Burman et al., Safely Expanding Research Access to Administrative Tax Data: Creating a Synthetic Public Use File and a Validation Server (2018); *see also The Synthetic Data Vault*, https://sdv.dev.
8 Valerie Chen, Valerio Pastro & Mariana Raykova, *Secure Computation for Machine Learning with SPDZ*, Cornell U. (Jan. 2, 2019), https://arxiv.org/pdf/1901.00329.pdf.
9 Louis J. M. Aslett et al., *A Review of Homomorphic Encryption and Software Tools for Encrypted Statistical Machine Learning*, Cornell U. (Aug. 26, 2015), https://arxiv.org/pdf/1508.06574.pdf.
10 *See* Hongyan Chang & Reza Shokri, *On the Privacy Risks of Algorithmic Fairness*, Cornell U. (Apr. 7, 2021), https://arxiv.org/pdf/2011.03731.pdf.
11 Ruggles et al., *Differential Privacy and Census Data: Implications for Social and Economic Research*, 109 Am. Econ. Ass'n Papers & Proceedings 403, 406 (2019).
12 In Computer Science literature, such algorithmic settings are often referred to as *hyperparameters*. For instance, *k* is a hyperparameter for *k*-anonymity. By setting *k* to different values (e.g., 5, 10, 100), practitioners can modulate the amount of anonymity afforded to records in the data. As we note however, the choice of hyperparameters controls both the privacy effected on a dataset as well as the fidelity of that data.
13 *See Differential Privacy for Census Data Explained*, Nat'l Conf. of State Legislatures (July 1, 2021), https://www.ncsl.org/research/redistricting/differential-privacy-for-census-data-explained.aspx; Hongyan Chang & Reza Shokri, *On the Privacy Risks of Algorithmic Fairness*, Cornell U. (Apr. 7, 2021), https://arxiv.org/pdf/2011.03731.pdf. Some scholars even find that the incorporation of differential privacy into machine learning algorithms can have disparate impact on underrepresented groups. *See* Eugene Bagdasaryan & Vitaly Shmatikov, *Differential Privacy Has Disparate Impact on Model Accuracy*, Cornell U. (Oct. 27, 2019), https://arxiv.org/pdf/1905.12101.pdf.
14 Steven Ruggles, Differential Privacy and Census Data: Implications for Social and Economic Research 17.
15 *Id.* at 18-19.
16 Nat'l Acad. of Sci., Innovations in Federal Statistics 86 (2017). The fragmented FSRDC review process is similar to the fragmented data access regime we discussed in Chapter Three.
17 *Special Sworn Researcher Program*, Bureau of Econ. Analysis, https://www.bea.gov/research/special-sworn-researcher-program (last updated July 23, 2021).
18 13 U.S.C. § 9.
19 The institutional form of the NRC is discussed in depth in Chapter Four.
20 NORC Data Enclave, NORC, https://www.norc.org/PDFs/BD-Brochures/2016/Data%20Enclave%20One%20Sheet.pdf.
21 *CMS Virtual Research Data Center (VRDC)*, Research Data Assistance Ctr., https://resdac.org/cms-virtual-research-data-center-vrdc.
22 *Request for Information (RFI) Seeking Stakeholder Input on the Need for an NIH Administrative Data Enclave*, Nat'l Inst. of Health (Mar. 1, 2019), https://grants.nih.gov/grants/guide/notice-files/NOT-OD-19-085.html.

23 *See FASEB Response to NIH Request for Information (RFI): Seeking Stakeholder Input on the Need for an NIH Administrative Data Enclave*, Fed'n of Am. Societies for Experimental Biology (2019), https://www.faseb.org/Portals/2/PDFs/opa/2019/FASEB_Response_Data_Enclave_RFI_NOT-OD-19-085.pdf; Am. Soc'y of Biochemistry & Molecular Biology (May 30, 2019), https://www.asbmb.org/getmedia/e3401ed5-3210-4ed2-a82a-7363cb86071d/ASBMB-Response-to-NIH-RFI-NOT-09-19-085.pdf.
24 *What We Do,* Cal. Pol'y Lab, https://www.capolicylab.org/what-we-do/.
25 *Id.*
26 *CPL Roadmap to Government Administrative Data in California*, Cal. Pol'y Lab, https://www.capolicylab.org/data-resources/california-data-road-map/.
27 Interview with Evan White, Executive Director, California Policy Lab (Apr. 29, 2021).
28 *Id.*
29 *Id.*; *see, e.g., Policy Evaluation and Research Linkage Initiative (PERLI)*, Cal. Pol'y Lab, https://www.capolicylab.org/data-resources/perli/; *University of California Consumer Credit Panel*, Cal. Pol'y Lab, https://www.capolicylab.org/data-resources/university-of-california-consumer-credit-panel/.
30 Interview with Evan White, *supra* note 27.
31 *Id.*
32 *See, e.g., Life Course Dataset*, Cal. Pol'y Lab, https://www.capolicylab.org/life-course-dataset/.
33 *See CPL Roadmap to Government Administrative Data in California*, *supra* note 26.
34 We note that it is possible that the organizational form could affect the authority of NRC staff to speak to the legality of data transfers.

Chapter 7

1 Christopher Whyte, *Deepfake News: AI-Enabled Disinformation as a Multi-Level Public Policy Challenge*, 5 J. Cyber Pol'y 199 (2020); Don Fallis, *What Is Disinformation?*, 63 Libr. Trends 601 (2015).
2 Mary L. Gray & Siddharth Suri, Ghost Work: How to Stop Silicon Valley From Building a New Global Underclass (2019); *Science Must Examine the Future of Work*, Nature (Oct. 19, 2017), https://www.nature.com/articles/550301b.
3 David Danks & Alex John London, *Algorithmic Bias in Autonomous Systems*, 26 Int'l Joint Conf. on Artificial Intelligence 4691 (2017); Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 Proceeding of Machine Learning Res. 1 (2018); Ben Hutchinson et al., *Unintended Machine Learning Biases as Social Barriers for Persons with Disabilities*, 125 ACM SIGACCESS Accessibility & Computing 1 (2020).
4 Oscar H. Gandy Jr., The Panoptic Sort: A Political Economy of Personal Information (1993); Virginia Eubanks, Automating Inequality (2018); Rashida Richardson, *Racial Segregation and the Data-Driven Society: How Our Failure to Reckon with Root Causes Perpetuates Separate and Unequal Realities*, 36 Berkeley Tech. L. J. 101 (2021).
5 For approaches to improve machine learning practices, see Timnit Gebru et al., *Datasheets for Datasets*, Cornell U. (Mar. 19, 2020), https://arxiv.org/pdf/1803.09010.pdf; Margaret Mitchell et al., *Model Cards for Model Reporting*, 2019 Proceedings ACM Conf. on Fairness, Accountability & Transparency 220 (2019); Kenneth Holstein et al., *Improving Fairness in Machine Learning Systems: What do Industry Practitioners Need?*, 2019 CHI Conf. on Hum. Factors in Computing Sys. 1 (2019); Michael A Madaio et al., *Co-Designing Checklists to Understand Organizational Challenges and Opportunities Around Fairness in AI*, 2020 CHI Conf. on Hum. Factors in Computing Sys. 318 (2019). The literature on AI's societal impacts and fairness, accountability, and transparency of AI is vast, but see Michael Kearns & Aaron Roth, The Ethical Algorithm: The Science of Socially Aware Algorithm Design (2019); Eubanks, *supra* note 4; Solon Barocas, Moritz Hardt & Arvind Narayanan, Fairness and Machine Learning (2019); Cathy O'Neil, Weapons of Math Destruction (2016).
6 45 C.F.R §§ 46.101-124.
7 J. Britt Holbrook & Robert Frodeman, *Peer Review and the Ex Ante Assessment of Societal Impacts*, 20 Res. Evaluation 239 (2011).
8 *Id.*
9 *Institutional Review Boards (IRBs) and Protection of Human Subjects in Clinical Trials*, U.S. Food & Drug Admin., https://www.fda.gov/about-fda/center-drug-evaluation-and-research-cder/institutional-review-boards-irbs-and-protection-human-subjects-clinical-trials (last updated Sept. 11, 2019).
10 There are crucial questions with regards to consent even with data considered "publicly" available. *See generally* Casey Fiesler & Nicholas Proferes, *"Participant" Perceptions of Twitter Research Ethics*, 4 Social Media + Society 1 (2018); Sarah Gilbert, Jessica Vitak & Katie Shilton, *Measuring Americans' Comfort with Research Uses of Their Social Media Data*, 7 Social Media + Society 1 (2021).
11 Sara R. Jordan, Future of Privacy Forum, Designing an Artificial Intelligence Research Review Committee (2019), https://fpf.org/wp-content/uploads/2019/10/DesigningAIResearchReviewCommittee.pdf.
12 Agatta Feretti et al., *Ethics Review of Big Data Research: What Should Stay and What Should Be Reformed?*, 22 BMC Medical Ethics 1, 6 (2021); Kathryn M. Porter et al., *The Emergence of Clinical Research Ethics Consultation: Insights from a National Collaborative*, 2018 Am. J. Bioethics 39 (2018).
13 Feretti et al., *supra* note 12.
14 *See, e.g.,* Mark Diaz et al., *Addressing Age-Related Bias in Sentiment Analysis*, 2018 Proceedings CHI Conf. on Hum. Factors in Computing Sys. 1 (2018); Buolamwini & Gebru, *supra* note 3.
15 *See, e.g.,* Timnit Gebru et al., *Datasheets for Datasets*, Cornell U. (Mar. 19, 2020), https://arxiv.org/pdf/1803.09010.pdf; Margaret Mitchell et al., *Model Cards for Model Reporting*, 2019 Proceedings ACM Conf. on Fairness, Accountability & Transparency 220 (2019); Emily M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, 2021 Proceedings ACM Conf. on Fairness, Accountability & Transparency 610 (2021); Christo Wilson et al., *Building and Auditing Fair Algorithms: A Case Study in Candidate Screening*, 2021 Proceedings ACM Conf. on Fairness, Accountability & Transparency 666 (2021); Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. Pa. L. Rev. Online 189 (2017).
16 *Phase II: Proposal Review and Processing*, Nat'l Sci. Found., https://www.nsf.gov/bfa/dias/policy/merit_review/phase2.jsp#select.
17 Harvey A. Averch, *Criteria for Evaluating Research Projects and Portfolios*, *in* Evaluating R&D Impacts: Methods and Practice 263 (1993).
18 Nat'l Security Comm'n on Artificial Intelligence, Final Report 141-54 (2021).
19 *History and Mission*, U.S. Privacy & Civil Liberties Oversight Bd., https://www.pclob.gov/About/HistoryMission.
20 AI in Counterterrorism Oversight Enhancement Act of 2021, H.R. 4469, 117th Cong. (2021).
21 In instances where a researcher is using data obtained from one of the agencies that falls under ORI's oversight, it may make sense to have ORI adju-

dicate those cases directly. For more information about the ORI, see *ORI*, Office Of Research Integrity, https://ori.hhs.gov/.
22 Michael S. Bernstein et al., *ESR: Ethics and Society Review of Artificial Intelligence Research*, Cornell U. (July 9, 2021), https://arxiv.org/pdf/2106.11521.pdf.
23 Nat'l Sci. Found., *Broader Impacts*, https://www.nsf.gov/od/oia/special/broaderimpacts/.
24 *See, e.g., Notice of Special Interest: Administrative Supplement for Research and Capacity Building Efforts Related to Bioethical Issues*, Nat'l Inst. of Health (Nov. 17, 2020), https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-020.html; *Notice of Special Interest: Administrative Supplement for Research on Bioethical Issues*, Nat'l Inst. of Health (Dec. 30, 2019), https://grants.nih.gov/grants/guide/notice-files/NOT-OD-20-038.html; *see also* Courtenay R. Bruce et al., *An Embedded Model for Ethics Consultation: Characteristics, Outcomes, and Challenges*, 5 AJOB Empirical Bioethics 8 (2014); Sharon Begley, *In a Lab Pushing the Boundaries of Biology, an Embedded Ethicist Keeps Scientists in Check*, Stat (Feb. 23, 2017), https://www.statnews.com/2017/02/23/bioethics-harvard-george-church/. Private foundations also promote the use of embedded bioethicists. *See, e.g., Making a Difference Request for Proposals – Fall 2021*, The Greenwall Found. (2021), https://greenwall.org/making-a-difference-grants/request-for-proposals-MAD-fall-2021.

Chapter 8

1 Paul Cichonski et al., Computer Security Incident Handling Guide (2012).
2 Putative attacks could include the deployment of ransomware, phishing schemes, gaining root access (the highest level of privilege available which gives users access to all commands and files by default), exposure of secret credentials, data poisoning, data exfiltration, as well as other types of unauthorized network intrusions.
3 Karen Hao, *AI Consumes a Lot of Energy. Hackers Could Make it Consume More*, MIT Tech. R. (May 6, 2021), https://www.technologyreview.com/2021/05/06/1024654/ai-energy-hack-adversarial-attack/.
4 Catalin Cimpanu, *Vast Majority of Cyber-Attacks on Cloud Servers Aim to Mine Cryptocurrency*, ZDNet (Sept. 14, 2020), https://www.zdnet.com/article/vast-majority-of-cyber-attacks-on-cloud-servers-aim-to-mine-cryptocurrency/.
5 We note that the NRC will likely need to comply with data specific security regulations as well. For instance, medical data security will need to comply with HIPAA, and financial data will need to comply with The Gramm-Leach-Bliley Act.
6 Ray Dunham, *FISMA Compliance: Security Standards & Guidelines Overview*, Linford & Co. (Nov. 29, 2017), https://linfordco.com/blog/fisma-compliance/.
7 Amy J. Frontz, Review of the Department of Health and Human Services Compliance with the Federal Information Security Modernization Act of 2014 for Fiscal Year 2020 (2021).
8 U.S. Senate Comm. on Homeland Security & Governmental Affairs, Federal Cybersecurity: America's Data at Risk 18 (2019).
9 *Federal Information Security Modernization Act (FISMA) Background*, Nat'l Inst. Standards & Tech., https://csrc.nist.gov/projects/risk-management/fisma-background (last updated Aug. 4, 2021).
10 Dunham, *supra* note 6.
11 U.S. Senate Comm. on Homeland Security & Governmental Affairs, *supra* note 8, at 19.
12 *Id.* at 18.
13 *Id.* at 19.
14 *Id.* at 20.
15 Kevin Stine, et al., Guide for Mapping Types of Information and Information Systems to Security Categories (2008). Specifically, FISMA defines compliance in terms of three levels: low impact, moderate impact, and high impact. Low impact indicates that the loss of confidentiality, integrity, or availability of the system will have a limited adverse effect, while high impact indicates that such losses will have severe or catastrophic effects. *See* Sarah Harvey, *3 FISMA Compliance Levels: Low, Moderate, High*, KirkpatrickPrice (Apr. 24, 2020), https://kirkpatrickprice.com/blog/fisma-compliance-levels-low-moderate-high/.
16 Nat'l Inst. Standards & Tech., Security and Privacy Controls for Information Systems and Organizations (2020).
17 Marianne Swanson et al., Guide for Developing Security Plans for Federal Information Systems (2006).
18 Michael McLaughlin, *Reforming FedRAMP: A Guide to Improving the Federal Procurement and Risk Management of Cloud Services*, Info. Tech. & Innovation Found. (June 15, 2020), https://itif.org/publications/2020/06/15/reforming-fedramp-guide-improving-federal-procurement-and-risk-management.
19 *Program Basics*, FedRAMP, https://www.fedramp.gov/program-basics/; *see also* Steven VanRoekel, Security Authorization of Information Systems in Cloud Computing Environments (2011).
20 *FISMA vs. FedRAMP and NIST: Making Sense of Government Compliance Standards*, Foresite, https://foresite.com/fisma-vs-fedramp-and-nist-making-sense-of-government-compliance-standards/. However, we note that FedRAMP approval is exempted for certain types of cloud models: (i) where the cloud is private to the agency, (ii) where the cloud is physically located within a Federal facility, (iii) where the agency is not providing cloud services from the cloud-based information system to any external entities. *See* VanRoekel, *supra* note 19.
21 FedRAMP, FedRAMP Security Assessment Framework 5 (2017).
22 Doina Chiacu, *White House Warns Companies to Step Up Cybersecurity: 'We Can't Do it Alone'*, Reuters (June 3, 2021), https://www.reuters.com/technology/white-house-warns-companies-step-up-cybersecurity-2021-06-03/; *see also Significant Cyber Incidents*, Ctr. Strategic & Int'l Studies, https://www.csis.org/programs/strategic-technologies-program/significant-cyber-incidents (last visited Aug. 19, 2021).
23 U.S. Senate Comm. Homeland Security & Governmental Affairs, *supra* note 8, at 5.
24 *Id.* at 6.
25 Frontz, *supra* note 7.
26 Jonathan Reiber & Matt Glenn, *The U.S. Government Needs to Overhaul Cybersecurity. Here's How.*, Lawfare (Apr. 9, 2021), https://www.lawfareblog.com/us-government-needs-overhaul-cybersecurity-heres-how.
27 Nat'l Security Agency, Embracing a Zero Trust Security Model (2021).
28 McLaughlin, *supra* note 18.
29 *Id.*
30 *Id.*
31 Exec. Order No. 14,028, 86 Fed. Reg. 26633 (May 17, 2021).
32 U.S. Office of Mgmt. & Budget, Moving the U.S. Government Towards Zero Trust Cybersecurity Principles (2021).

33 *See, e.g.*, David Kushner, *The Real Story of Stuxnet*, IEEE Spectrum (Feb. 26, 2013), https://spectrum.ieee.org/the-real-story-of-stuxnet.
34 HTTP is the protocol at the highest level of abstraction targeting the application layer, and its secure variant HTTPS additionally encrypts the data using an encryption protocol. Without encryption, HTTP is insecure and should not be used. The encryption protocol in original use was SSL but this has since been deprecated in the realm of network security in favor of its newer version, TLS. Both SSL and TLS rely on public key certificates signed by a trusted certificate authority. When these certificates have expired, the websites providing them can no longer necessarily be trusted. Although these measures have their own limitations, not adopting them can only be less secure.
35 *See, e.g., Azure Confidential Computing*, Microsoft, https://azure.microsoft.com/en-ca/solutions/confidential-compute/; Nataraj Nagaratnam, *Confidential Computing*, IBM (Oct. 16, 2020),
https://www.ibm.com/cloud/learn/confidential-computing; *Confidential Computing*, Google Cloud, https://cloud.google.com/confidential-computing.
36 David Archer et al., *From Keys to Databases—Real-World Applications of Secure Multi-Party Computation*, 61 Computer J. 1749 (2018).
37 Amit Elazari Bar On, *We Need Bug Bounties for Bad Algorithms*, Motherboard (May 3, 2018) https://www.vice.com/en/article/8xkyj3/we-need-bug-bounties-for-bad-algorithms.

Chapter 9

1 Importantly, this chapter discusses the extent to which researchers should be required to share their research outputs, *not* the extent to which researchers should be required to share their private data. The latter was discussed in Chapter Three.
2 Dan Robitzski, *AI Researchers Are Boycotting A New Journal Because It's Not Open Access*, Futurism (May 3, 2018), https://futurism.com/artificial-intelligence-journal-boycot-open-access.
3 Mikio L. Braun & Cheng Soon Ong, Open Science in Machine Learning (2014).
4 Since researchers using the NRC are not "contractors" under FAR/DFARS, and since evidence is lacking on the value of Other Transactions to AI researchers, we do not cover FAR/DFARS and Other Transactions in this section.
5 Under the Bayh-Dole Act, Aa "federal funding agreement" is defined as "any contract, grant, or cooperative agreement entered into between any Federal agency, other than the Tennessee Valley Authority, and any contractor for the performance of experimental, developmental, or research work funded in whole or in part by the Federal Government." 35 U.S.C. § 201.
6 35 U.S.C. § 202.
7 35 U.S.C. § 203.
8 *See, e.g.*, Mark A. Lemley & Julie E. Cohen, *Patent Scope and Innovation in the Software Industry*, 89 Cal. L. Rev. 1 (2001); Mark A. Lemley, *Software Patents and the Return of Functional Claiming*, 2013 Wis. L. Rev. 905 (2013).
9 Jeremy Gillula & Daniel Nazer, *Stupid Patent of the Month: Will Patents Slow Artificial Intelligence?,* Elec. Frontier Found. (Sept. 29, 2017), https://www.eff.org/deeplinks/2017/09/stupid-patent-month-will-patents-slow-artificial-intelligence.
10 U.S. Patent & Trademark Off., Inventing AI: Tracing the Diffusion of Artificial Intelligence with Patents 2 (2020).
11 *See, e.g.,* Mike James, *Google Files AI Patents*, I Programmer (July 8, 2015), https://www.i-programmer.info/news/105-artificial-intelligence/8765-google-files-ai-patents.html. This is especially problematic because companies represent 26 out of the top 30 AI patent applicants worldwide, while only four are universities or public research organizations. World Intell. Prop. Org., Artificial Intelligence 7 (2019).
12 Lisa Ouellette & Rebecca Weires, *University Patenting: Is Private Law Serving Public Values?*, 2019 Mich. St. L. Rev. 1329 (2019).
13 *Id.* at 1331; *see also* Arti Kaur Rai, *Regulating Scientific Research: Intellectual Property Rights and the Norms of Science*, 94 Nw. U. L. Rev. 77, 136 (1999).
14 *See* Brian J. Love, *Do University Patents Pay Off? Evidence From a Survey of University Inventors in Computer Science and Electrical Engineering*, 16 Yale J. L & Tech. 285 (2014).
15 *See id.* at 286.
16 *See, e.g., Tech Transfer FAQ*, U. Mich., https://techtransfer.umich.edu/for-inventors/resources/inventor-faq/ ("We carefully review the commercial potential for an invention before investing in the patent process. However, because the need for commencing a patent filing usually precedes finding a licensee, we look for creative and cost-effective ways to seek early protections for as many promising inventions as possible"); *What is Technology Transfer*, Princeton U., https://patents.princeton.edu/about-us/what-technology-transfer ("[T]echnologies and everyday products are possible because of technology transfer . . . Because the discoveries emerging from university research tend to be early-stage, high-risk inventions, successful university technology transfer transactions require a patent system that protects such innovations.").
17 The Uniform Guidance for intellectual property is laid out in 2 C.F.R. § 200.315.
18 *Uniform Administrative Requirements, Cost Principles, and Audit Requirements for Federal Awards*, Grants.gov, https://www.grants.gov/learn-grants/grant-policies/omb-uniform-guidance-2014.html (last visited Aug. 27, 2021).
19 *See Key Sections of the Uniform Guidance*, AICPA.org, https://www.aicpa.org/interestareas/governmentalauditquality/resources/singleaudit/uniformguidanceforfederalrewards/key-sections-uniform-guidance.html.
20 2 C.F.R. § 200.315. A "federal award" under the Uniform Guidance includes, among other things, "the federal financial assistance that a recipient receives directly from a Federal awarding agency or indirectly from a pass-through entity;" or "the cost-reimbursement contract under the Federal Acquisition Regulations;" or a "grant agreement, cooperative agreement, [or] other agreement [for federal financial assistance]." 2 C.F.R. § 200.1.
21 2 C.F.R. §§ 200.315(b), (c). These provisions specify that the government merely "reserves" its "right" to copyright and data rights over research produced under the federal award.
22 U.S. Copyright Office, Compendium of U.S. Copyright Office Practices 35 (2021, 3d ed.).
23 Wil Michiels, *How Do You Protect Your Machine Learning Investment?*, EETimes (Mar. 31, 2020), https://www.eetimes.com/how-do-you-protect-your-machine-learning-investment-part-ii/.
24 *See, e.g.,* Tabrez Y. Ebrahim, *Data-Centric Technologies: Patent and Copyright Doctrinal Disruptions*, 43 Nova L. Rev. 287, 304; Daryl Lim, *AI & IP: Innovation & Creativity in an Age of Accelerated Change*, 52 Akron L. Rev. 813, 835 (2018)
25 2 C.F.R. § 200.315(b).
26 *Id.*
27 For a comprehensive report on how artificial intelligence is used in various government agencies, see David Freeman Engstrom, Daniel E. Ho, Catherine M. Sharkey & Mariano-Florentino Cuéllar, Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies (2020).

28 *Jukebox*, OpenAI (Apr. 30, 2020), https://openai.com/blog/jukebox/.
29 *See, e.g.,* Shlomit Yanisky-Ravid, *Generating Rembrandt: Artificial Intelligence, Copyright, and Accountability in the 3A Era--the Human-Like Authors are Already Here--a New Model*, 27 Mich. St. L. Rev. 659 (2017); Kalin Hristov, *Artificial Intelligence and the Copyright Dilemma*, 57 J. Franklin Pierce Ctr. Intell. Prop. 431 (2017).
30 Kalin Hristov, *Artificial Intelligence and the Copyright Survey*, 16 J. Sci. Pol'y & Governance 1, 14-15 (2020).
31 *Id.* at 16.
32 *See What is Transfer Learning?*, TensorFlow (Mar. 31, 2020), https://www.tensorflow.org/js/tutorials/transfer/what_is_transfer_learning.
33 *See, e.g.*, Yunhui Guo et al., *SpotTune: Transfer Learning Through Adaptive Fine-Tuning*, Cornell U. (Nov. 2018), https://arxiv.org/pdf/1811.08737.pdf.
34 2 C.F.R. § 200.315(d).
35 *See* Zhiqiang Wan, Yazhou Zhang & Haibo He, *Variational Autoencoder Based Synthetic Data Generation for Imbalanced Learning*, IEEE (2017).
36 *See* Noseong Park, Mahmoud Mohammadi & Kshitij Gorde, *Data Synthesis Based on Generative Adversarial Networks*, 11 Proc. VLDB Endowment 1071 (2018).
37 *See* Ron Bakker, Impact of Artificial Intelligence on IP Policy 12.
38 *See* Marta Duque Lizarralde, A Guideline to Artificial Intelligence, Machine Learning and Intellectual Property 4-7 (2020).
39 Steven M. Bellovin et al., *Privacy and Synthetic Datasets*, 22 Stan. Tech. L. Rev. 1, 2-3 (2019); *see also* Fida K. Dankar & Mahmoud Ibrahim, *Fake It Till You Make It: Guidelines for Effective Synthetic Data Generation*, 5 Applied Sci. 11 (2021); *but see* Theresa Stadler et al., *Synthetic Data – Anonymisation Groundhog Day*, Cornell U. (July 8, 2021), https://arxiv.org/pdf/2011.07018.pdf.
40 *See, e.g.,* Daniel S. Quintana, *A Synthetic Dataset Primer for the Biobehavioural Sciences to Promote Reproducibility and Hypothesis Generation*, 9 eLife 1 (2020).
41 Yuji Roh et al., *A Survey on Data Collection for Machine Learning*, Cornell U. (Aug. 12, 2019), https://arxiv.org/pdf/1811.03402.pdf.
42 *See, e.g.,* Hang Qiu et al., *Minimum Cost Active Labeling*, Cornell U. (June 24, 2020), https://arxiv.org/pdf/2006.13999.pdf; Eric Horvitz, *Machine Learning, Reasoning, and Intelligence in Daily Life: Directions and Challenges*, 18 Proceeding of the Conf. on Uncertainty in Artificial Intelligence 3 (2007).
43 Cognilytics Research, Data Engineering, Preparation, and Labeling for AI 2019 3 (2019).
44 *See* Wil Michiels, *How Do You Protect Your Machine Learning Investment?*, EETimes (Mar. 26, 2020), https://www.eetimes.com/how-do-you-protect-your-machine-learning-investment/. In fact, in the European Union, labeled datasets are awarded with database rights protections. Mauritz Kop, *Machine Learning & EU Data Sharing Practices*, Stan.-Vienna Transatlantic Tech. L. F.(Mar. 24, 2020), https://ttlfnews.wordpress.com/2020/03/24/machine-learning-eu-data-sharing-practices/.
45 *See, e.g.,* Niklas Fiedler et al., *ImageTagger: An Open Source Online Platform for Collaborative Image Labeling*, 11374 Lecture Notes in Computer Sci. 162 (2019).
46 *Id.* at 162.
47 Researchers may, for instance, use NRC data and compute resources to implement active learning strategies, procedures to manually label a subset of available data and infer the remaining labels automatically using a machine learning model. *See, e.g.,* Oscar Reyes et al., *Effective Active Learning Strategy for Multi-Label Learning*, 273 Neurocomputing 494 (2018). Similarly, researchers may augment existing public sector data with valuable labels.
48 *See, e.g.,* Pedro Saleiro et al., *Aequitas: A Bias and Fairness Audit Toolkit*, Cornell U. (Apr. 29, 2019), https://arxiv.org/pdf/1811.05577.pdf; Florian Tramèr et al., *FairTest: Discovering Unwarranted Associations in Data-Driven Applications*, Cornell U. (Aug. 16, 2019), https://arxiv.org/pdf/1510.02377.pdf.
49 While we do not discuss the idiosyncratic modifications to the Uniform Guidance that vary from agency-to-agency, we encourage the task force to assess these modifications if it decides to implement the NRC through a particular agency. If the NRC is administered through multiple agencies, the complex amalgam of agency-specific IP rules may increase the friction in using the NRC if researchers must context-switch from one set of regulations to the next depending on the funding agency.
50 2 C.F.R. § 2900.13. Previously, the Department of Labor explicitly required IP generated under a federal award to be licensed under a Creative Commons Attribution license, but this rule was changed in April 2021 to replace the proprietary term "Creative Commons Attribution license" with the industry-recognized standard "open license." 86 Fed. Reg. 22107 (Apr. 27, 2021).
51 *Dissemination and Sharing of Research Results - NSF Data Management Plan Requirements*, Nat'l Sci. Found., https://www.nsf.gov/bfa/dias/policy/dmp.jsp.
52 *See, e.g.,* Aidan Courtney et al., *Balancing Open Source Stem Cell Science with Commercialization*, Nature Biotechnology (Feb. 7, 2011), https://www.nature.com/articles/nbt.1773.
53 *See* Klint Finley, *When Open Source Software Comes with a Few Catches*, Wired (July 31, 2019), https://www.wired.com/story/when-open-source-software-comes-with-catches/; *Guide to Open Source Licenses*, Synopsys (Oct. 7, 2016), https://www.synopsys.com/blogs/software-security/open-source-licenses/.
54 *See* Daniel A. Almeida et. al, *Do Software Developers Understand Open Source Licenses?*, 25 IEEE Int'l Conf. on Program Comprehension 1 (2017) (finding that software developers "struggle[] when multiple [open-source] licenses [are] involved" and "lack the knowledge and understanding to tease apart license interactions across multiple situations.").
55 *See, e.g.,* Alexandra Theben et al., Challenges and Limits of an Open Source Approach to Artificial Intelligence 14 (2021); Stadler et al., *supra* note 39; Milad Nasr et al., *Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning*, Cornell U. (June 6, 2020), https://arxiv.org/abs/1812.00910.pdf.
56 Some universities have decided to eliminate classified research. *See, e.g.*, *At the Hands of Radicals*, Stan. Mag. (Jan. 2009), https://stanfordmag.org/contents/at-the-hands-of-the-radicals.
57 *See* Donald Kennedy, *Science and Secrecy*, 289 Sci. 724 (2000); Peter J. Westwick, *Secret Science: A Classified Community in the National Laboratories*, 38 Minerva 363 (2000).
58 *See* Braun & Ong, *supra* note 3; Sören Sonnenburg et al., *The Need for Open Source Software in Machine Learning*, 8 J. Machine Learning Res. 2443 (2007); *see also* Katie Malone & Richard Wolski, *Doing Data Science on the Shoulders of Giants: The Value of Open Source Software for the Data Science Community*, HDSR (May 31, 2020), https://hdsr.mitpress.mit.edu/pub/xsrt4zs2/release/4.
59 *See* Laura A. Heymann, *Overlapping Intellectual Property Doctrines: Election of Rights Versus Selection of Remedies*, 17 Stan. Tech. L. Rev. 239, 240 (2013); *Oracle Am. Inc. v. Google Inc.*, 750 F.3d 1339 (Fed. Cir. 2014) (accepting that software is both patentable and copyrightable).

60 Robert E. Thomas, *Debugging Software Patents: Increasing Innovation and Reducing Uncertainty in the Judicial Reform of Software Patent Law*, 25 Santa Clara Computer & High Tech. L.J. 191, 222-23 (2008).
61 *See, e.g.*, Joaquin Vanschorin et al., *OpenML: Networked Science in Machine Learning*, Cornell U. (Aug. 1, 2014), https://arxiv.org/pdf/1407.7722.pdf (developing a collaboration platform through which scientists can automatically share, organize and discuss machine learning experiments, data, and algorithms); *see also* Sarah O'Meara, *AI Researchers in China Want to Keep the Global-Sharing Culture Alive*, Nature (May 29, 2019), https://www.nature.com/articles/d41586-019-01681-x; Shuai Zhao et al., *Packaging and Sharing Machine Learning Models via the Acumos AI Open Platform*, 17 ICMLA (2018).
62 Jeanne C. Fromer, *Machines as the New Oompa-Loompas: Trade Secrecy, the Cloud, Machine Learning, and Automation*, 94 N.Y.U. L. Rev. 706, 712 (2019); Jordan R. Raffe et al., The Rising Importance of Trade Secret Protection for AI- Related Intellectual Property 1, 5-6 (2020); Jessica M. Meyers, *Artificial Intelligence and Trade Secrets*, Am. Bar Ass'n (Feb. 2019), https://www.americanbar.org/groups/intellectual_property_law/publications/landslide/2018-19/january-february/artificial-intelligence-trade-secrets-webinar/; *AIPLA Comments Regarding "Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation"*, Am. Intell. Prop. L. Ass'n (Jan. 10, 2020), https://www.uspto.gov/sites/default/files/documents/AIPLA_RFC-84-FR-58141.pdf.
63 Clark D. Asay, *Artificial Stupidity*, 61 Wm. & Mary L. Rev. 1187, 1197, 1241-42 (2020).
64 *See id.*; Am. Intell. Prop. L. Ass'n, *supra* note 62, at 16.
65 *See* Asay, *supra* note 63, at 1242.

Appendix

1 *Department of Energy Awards $425 Million for Next Generation Supercomputing Technologies*, Energy.gov (Nov. 14, 2014), https://www.energy.gov/articles/department-energy-awards-425-million-next-generation-supercomputing-technologies.
2 *Amazon EC2 P3 Instances*, Amazon, https://aws.amazon.com/ec2/instance-types/p3/ (last visited Sept. 9, 2021).
3 *CORAL Request for Proposal B604142*, Lawrence Livermore Nat'l Laboratory (2014), https://web.archive.org/web/20140816181824/ https://asc.llnl.gov/CORAL/. We note that we were not able to locate the final award documents, nor is Summit budgeted in sufficient detail to back out cost from the DOE budget statements. Our cost estimates here, however, are comparable to publicly reported estimates for the total cost of the Summit system.
4 This is based on a $30M maximum in the DOE Office of Science contract for non-recurring engineering (NRE) costs for the systems at Argonne National Laboratory and Oak Ridge National Laboratory.
5 This is based on the difference in the RFP terms between the inclusion of maintenance under the Lawrence Livermore National Laboratory system (with a maximum budget of $170M) and the exclusion of maintenance under the systems for the Oak Ridge National Laboratory and the Argonne National Laboratory (with a maximum budget for the build contract of $155M). This is likely an upper bound on maintenance, given that the difference reflects the combination of NRE and 5-year maintenance.
6 *See CORAL Price Schedule*, Lawrence Livermore Nat'l Laboratory (2014), https://web.archive.org/web/20140816181824/ https://asc.llnl.gov/CORAL/RFP_components/04_CORAL_Price_Schedule_ANL_ORNL_tabs.xlsx. We used 1.62% as the interest rate to calculate the cost over 60 months. It is the 5-year Treasury constant maturity rate on November 14, 2014, *see Selected Interest Rates (Daily) – H.15*, Fed. Res., https://www.federalreserve.gov/releases/H15/default.htm, when DOE announced the award of the HPC system, *see Department of Energy Awards $425 Million for Next Generation Supercomputing Technologies, supra* 1.
7 For instance, this estimate is in line with the cost of $200M reported by the *New York Times*. Steve Lohr, *Move Over, China: U.S. is Again Home to World's Speediest Supercomputer*, N.Y. Times (June 8, 2018), https://www.nytimes.com/2018/06/08/technology/supercomputer-china-us.html. Some reporting conflates the procurement of multiple systems that occurred contemporaneously.
8 Research shows that for training compute-intensive deep learning models, such as ResNet-101, the GPU utilization is around 70%. Jingoo Han et al., *A Quantitative Study of Deep Learning Training on Heterogeneous Supercomputers*, 2019 IEEE Conf. on Cluster Computing 1, 5 (2019). However, ResNet-50 has a GPU utilization of approximately 40%, *see id.*, and other accounts report that GPUs are utilized only 15-30% of the time, *see, e.g.*, Lukas Biewald, *Monitor and Improve GPU Usage for Training Deep Learning Models*, Towards Data Sci. (Mar. 27, 2019), https://towardsdatascience.com/measuring-actual-gpu-usage-for-deep-learning-training-e2bf3654bcfd; Janet Morss, *Giving Your Data Scientists a Boost with GPUaaS*, CIO (June 2, 2020), https://www.cio.com/article/3561090/giving-your-data-scientists-a-boost-with-gpuaas.html.
9 Compute Canada, Cloud Computing for Researchers 1 (2016), https://www.computecanada.ca/wp-content/uploads/2015/02/CloudStrategy2016-2019-forresearchersEXTERNAL-1.pdf.
10 Jennifer Shkabatur, *The Global Commons of Data*, 22 Stan. Tech. L.R. 407, 407-09 (2019).
11 Benjamin Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 Colum. J.L. & Arts 61 (2017).
12 *Id.*
13 *See Protecting What We Love About the Internet: Our Efforts to Stop Online Piracy*, Google Pub. Pol'y Blog (Nov. 7, 2019), https://www.blog.google/outreach-initiatives/public-policy/protecting-what-we-love-about-internet-our-efforts-stop-online-piracy/.
14 *See* Jennifer M. Urban, Joe Karaganis & Brianna M. Schofield, Notice & Takedown in Everyday Practice 39 (2017) (illustrating the difficulty that online service providers face in manually evaluating a large volume of data for potential infringement; for example, one online service provider explained that "out of fear of failing to remove infringing material, and motivated by the threat of statutory damages, its staff will take "six passes to try to find the [identified content]."); *see also* Letter from Thom Tillis, Marsha Blackburn, Christopher A. Coons, Dianne Feinstein et. al, to Sundar Pichai, Chief Executive Officer, Google Inc. (Sept. 3, 2019), https://www.ipwatchdog.com/wp-content/uploads/2019/09/9.3-Content-ID-Ltr.pdf ("We have heard from copyright holders who have been denied access to Content ID tools, and as a result, are at a significant disadvantage to prevent repeated uploading of content that they have previously identified as infringing. They are left with the choice of spending hours each week seeking out and sending notices about the same copyrighted works, or allowing their intellectual property to be misappropriated.").
15 *See* Google, How Google Fights Piracy 6 (2016). To illustrate the costs of implementing Content ID on a large-scale platform, Google announced in a report in 2016 that YouTube had invested more than $60 million in Content ID.
16 *See* Sobel, *supra* note 11, at 66-79.
17 *See Authors Guild v. Google Inc.*, 804 F.3d 202 (2d Cir. 2015).
18 *Id*.
19 *Id.* at 216-17.
20 Matthew Stewart, *The Most Important Court Decision For Data Science and Machine Learning*, Towards Data Sci. (Oct. 31, 2019), https://towardsdata-

science.com/the-most-important-supreme-court-decision-for-data-science-and-machine-learning-44cfc1c1bcaf.
21 *See, e.g.*, James Grimmelmann, *Copyright for Literate Robots*, 101 Iowa L. Rev. 657, 661; Sobel, *supra* note 11, at 51-57.
22 *See* Sobel, *supra* note 11, at 57.
23 *See* Anna I. Krylov et. al. *What is the Price of Open Source Software?* 6 J. Physical Chemistry Letters 2751, 2753 (2015) (explaining that budding researchers considering commercialization may be particularly concerned about what licenses are available, since a "strictly open-source environment may furthermore disincentivize young researchers to make new code available right away, lest their ability to publish papers be short-circuited by a more senior researcher with an army of postdocs poised to take advantage of any new code.").
24 *See, e.g., A Data Scientist's Guide to Open-Source Licensing,* Towards Data Sci. (Nov. 4, 2018), https://towardsdatascience.com/a-data-scientists-guide-to-open-source-licensing-c70d5fe42079; *Choose an Open-Source License*, https://choosealicense.com.
25 *Licensing a Repository,* GitHub, https://docs.github.com/en/github/creating-cloning-and-archiving-repositories/licensing-a-repository.
26 *What is the Most Appropriate Licence for My Data?*, FigShare, https://help.figshare.com/article/what-is-the-most-appropriate-licence-for-my-data.
27 *See Developer Agreement*, Twitter (Mar. 10, 2020), https://developer.twitter.com/en/developer-terms/agreement; *Non-commercial Use of the Twitter API*, Twitter, https://developer.twitter.com/en/developer-terms/commercial-terms.
28 *See* Daniel A. Almeida et. al, *Do Software Developers Understand Open Source Licenses?*, 25 IEEE Int'l Conf. on Program Comprehension 1 (2017).
29 *Id.* at 9.
30 Alexandra Kohn & Jessica Lange, *Confused About Copyright? Assessing Researchers' Comprehension of Copyright Transfer Agreements,* 6 J. Librarianship & Scholarly Commc'n. 1, 9 (2018).
31 *See* Will Frass, Jo Cross & Victoria Gardner, Taylor & Francis Open Access Survey June 2014 15 (2014). Note that lack of IP literacy could act as an additional deterrent to uploaders. The Taylor and Francis Open Access Survey of 2014 found that "63% of respondents indicated a lack of understanding of publisher policy as an important or very important factor in failing to deposit an article in an IR [Institutional Repository]." *Id.*
32 *Dataverse Community Norms*, Harv. Dataverse, https://dataverse.org/best-practices/dataverse-community-norms.
33 *Copyright and License Policy*, FigShare, https://help.figshare.com/article/copyright-and-license-policy.
34 Australian Data Research Commons, Research Data Rights Managing Guide 6 (2019).
35 *See Harvard Dataverse General Terms of Use*, Harv. Dataverse (2021), https://dataverse.org/best-practices/harvard-dataverse-general-terms-use.
36 Stan. U. Inst. of Human-Centered Artificial Intelligence, Artificial Intelligence Index Report 2021 125-34 (2021).
37 Thilo Hagendorff, *The Ethics of AI Ethics: An Evaluation of Guidelines*, 30 Minds & Machines 99 (2020).
38 Andrew D. Selbst, *An Institutional View of Algorithmic Impact Assessments*, 35 Harv. J.L. & Tech. 1, 66 (forthcoming 2021).
39 Brent Mittlestadt, *Principles Alone Cannot Guarantee Ethical AI*, 1 Nature Mach. Intelligence 501 (2019).
40 *DOD Adopts Ethical Principles for Artificial Intelligence*, U.S. Dep't Defense (Feb. 24, 2020), https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/.
41 President's Mgmt. Agenda, Federal Data Strategy: Data Ethics Framework (2020).
42 *Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities*, U.S. Gov't Accountability Office (June 30, 2021), https://www.gao.gov/products/gao-21-519sp.
43 *Principles of Artificial Intelligence Ethics for the Intelligence Community*, Office of the Director of Nat'l Intelligence, https://www.odni.gov/index.php/features/2763-principles-of-artificial-intelligence-ethics-for-the-intelligence-community.
44 *Key Considerations for Responsible Development and Fielding of Artificial Intelligence*, Nat'l Security Comm'n Artificial Intelligence (2021), https://www.nscai.gov/key-considerations/.
45 *Recommended Practices*, Nat'l Security Comm'n Artificial Intelligence, https://www.nscai.gov/wp-content/uploads/2021/01/Key-Considerations-Supporting-Visuals.pdf.
46 Defense Innovation Bd., AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense (2019).

HAI

Stanford University
Human-Centered
Artificial Intelligence

Stanford
Law School