UNIVERSITY of VIRGINIA

SCHOOL of DATA SCIENCE

School of Data Science
1919 Ivy Road, Room 336
Charlottesville, VA 22903

P 434.982-2600
reia@virginia.edu
datascience.virginia.edu

July 16, 2024

U.S. Department of Commerce
1401 Constitution Ave NW
Washington, DC 20230
United States

Re: AI and Open Government Data Assets Request for Information

The researchers in the School of Data Science (SDS) at the University of Virginia (UVA) appreciate the opportunity to submit comments to the U.S. Department of Commerce (Commerce) Request for Information (RFI) on AI and Open Government Data Assets (89 FR 27411).

Introduction

Open data is a key tool for government transparency and accountability. Thus, the improvement of processes to create, curate, distribute and maintain governmental open assets is crucial to safeguarding democracy. Over the past years, we saw the rapid dissemination of artificial intelligence (AI) systems worldwide, especially those based on large language models (LLMs) and other forms of generative AI that use deep learning models. The importance of open data assets to train such models has led different organizations to seek ways to make their data available to stakeholders pursuing innovative solutions for long-lasting, complex societal problems.

However, besides the potential for technological innovation, there are several concerns, harms and ethical issues that arise with making data available and using them to train AI models. We must move beyond the idea that simply making various datasets available will ensure transparency; instead, we should seek to understand the needs and demands of individuals and organizations,[1] to truly use open data assets for the public interest. Having clear goals, understanding the environmental impact context, addressing social implications of opening data, investing in a robust infrastructure and in data governance to make data available is a starting point of a long journey. The reason for such a complex process is because "we often think about open data as being technical or part of a country's digital transformation. But really, open data is about people, their problems, and giving them the power to solve them."[2] When planning to prepare open government data assets

[1] Erna Ruijer, Stephan Grimmelikhuijsen, and Albert Meijer. "Open Data for Democracy: Developing a Theoretical Framework for Open Data Use." *Government Information Quarterly* 34, no. 1 (January 2017): 45–52. https://doi.org/10.1016/j.giq.2017.01.001.
[2] Ana Brandusescu, "Open Data Is about People, Not Just Innovation." SciDev.Net, June 2017. https://www.scidev.net/global/opinions/open-data-people-innovation/.

for the use of AI, the human and environmental aspects of this process should not be an afterthought.

According to the 2022 Global Data Barometer report,[3] it is possible to shape data for the public good, but we have a long way to go. The world saw quick changes brought by the popularization of AI – systems that need massive volumes of data and resources – while open data around the globe have not advanced at the same speed. The U.S. scores relatively low in terms of data availability and use and impact,[4] showing that as AI becomes the center of attention in this process, there are structural issues in open government data that need to be addressed.

This document will focus on issues and considerations in (1) data ethics and digital rights, (2) data dissemination standards, (3) data integrity and quality, (4) partnerships and (5) climate action. It is divided based on the RFI's sections, featuring specific questions for that section that are being answered.

1. Data ethics and digital rights

Responding to Q1 and Q2.

The are already data ethics tenets available to guide Commerce, such as the recent "Data Ethics Framework," part of the U.S. Federal Data Strategy.[5] These principles state that data leaders in government must be in compliance with current regulations and policies, as well as respect privacy and confidentiality, listen to communities, be accountable, improve transparency and keep up to date with rapid developments in the field of data science. Here, we offer other recommendations that draw from and complement the Data Ethics Framework.

The Open Definition[6] by the Open Knowledge Foundation sets out principles for open data globally: open data "means anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness)." Building upon this, data must be digitally available, fully machine-readable, free or at no more than the cost of reproduction and without any restrictive licenses. Beyond these principles, practice also shapes if, how, when and why open government data will be available to individuals and organizations.

Furthermore, to reach equitable outcomes, Commerce should work to understand the limits of AI, starting with data collection. Public data and official numbers are powerful, and Commerce has a responsibility when releasing them and allowing AI models to be trained on them: "official data sets come out of offices and are imbued with the authority of those in power. They are offered publicly with the expectation that when given, they will be taken. They are a shared basis for the building of arguments and algorithms".[7] Understanding the impact of this authority is important, as overlooking it could exacerbate the possible biases, misrepresentations and limitations that may arise in government data. Specifically, without intentional steps to increase equity and access to the

---

[3] Global Data Barometer (2022). First Edition Report – Global Data Barometer. IL- DA. DOI: https://doi.org/10.5281/zenodo.6488349
[4] *Ibidem.*
[5] General Services Administration, *Federal Data Strategy Data Ethics Framework,* 2020, https://resources.data.gov/assets/documents/fds-data-ethics-framework.pdf.
[6] Open Knowledge, "The Open Definition." Open Content, n.d. https://opendefinition.org/.
[7] Dan Bouk, Kevin Ackermann, and danah boyd. "A Primer on Powerful Numbers: Selected Readings in the Social Study of Public Data and Official Numbers." Data & Society, March 2022. https://datasociety.net/library/a-primer-on-powerful-numbers-selected-readings-in-the-social-study-of-public-data-and-official-numbers/.

benefits of open data "a primary impact of 'open data' may be to further empower and enrich the already empowered and the well provided for".[8]

For instance, if data on certain groups, or regarding certain communities, is systematically under- or over-reported this will create inaccurate and inequitable models and outputs.[9] These biases in data can arise in various and unexpected ways. Take the 2020 Census's use of differential privacy as a means of protecting individual personal information, for instance; researchers found that the data representing a state's small Hispanic population was "far more impacted" by this Census privacy protection technique than the non-Hispanic population, making the data representing this population more inaccurate, which "present[s] a problem for researchers trying to study small, marginalized populations".[10] Additionally, the 2020 Census' use of synthetic data to fill in missing data for homes "that had not filled out the survey, had not been home when visited by an enumerator, and for which a "proxy enumeration" was not collected from a neighbor" likely also distorted the data in a patterned way making assumptions about homes where all adults are working and are not home to answer questions.[11]

Along with incorrect and assumed data, there is the possibility of governmental datasets missing data, or under collecting data on certain communities and phenomena. For instance, along with collecting information on chemicals and pollutants; generally, the Environmental Protection Agency (EPA) analyzes the risks of chemicals and pollutants to higher risks groups such as pregnant women and children; however, they have yet to look at the impacts of these chemicals on those in "fence-line" communities, meaning those living near industrial pollution and affected by a multitude of different chemicals at a time.[12] Further, a "100-mile stretch of the Mississippi River" often known as Cancer Alley, has only one Particulate Matter (PM2.5) monitor.[13] These limitations to governmental data are only known because communities and other organizations take the time to understand and report on the problems, and advocate for solutions. As such, there are likely many other limitations to governmental data that simply go unrecognized.

While the faults of datasets pose problems generally, there are certain risks that arise when they are used to train AI specifically. Because generative AI can appear to have trustworthy knowledge, and reliable understandings of the world, the data reports can become naturalized, or seen as objective, despite the complexity of all the choices that went into how the data was collected, and the opacity in possible biases and error that exist within it.[14] For example, a model built using EPA data discussed above on the harms of certain chemicals, would be unable to identify risks posed by exposure to multiple chemicals at the same time because the relevant data does not exist; that does not mean, however, that such harms do not exist, yet the model might lead us to believe as such. A shift towards generative AI to understand social, economic, or environmental problems should not

---

[8] Michael B. Gurstein, "Open Data: Empowering the Empowered or Effective Data Use for Everyone?," *First Monday*, January 23, 2011, https://doi.org/10.5210/fm.v16i2.3316.
[9] Jonas Lerman, "Big Data and Its Exclusions," *SSRN Electronic Journal*, 2013, https://doi.org/10.2139/ssrn.2293765.
[10] Simson Garfinkel, "Differential Privacy and the 2020 US Census," *MIT Case Studies in Social and Ethical Responsibilities of Computing*, no. Winter 2022 (January 27, 2022), https://doi.org/10.21428/2c646de5.7ec6ab93.
[11] *Ibidem.*
[12] Pat Rizzuto, "EPA Denies Pledge to 'Cancer Alley' Communities on Chemical Risks," 2020, https://news.bloomberglaw.com/environment-and-energy/epa-denies-pledge-to-cancer-alley-communities-on-chemical-risks.
[13] Kimberly Terrell and Wesley James, "Air Pollution and COVID-19: A Double Whammy for African American and Impoverished Communities in Cancer Alley," 2020, https://law.tulane.edu/sites/default/files/Files/Terrell%20-%20COVID-19%20-%20PM%202.5%20Louisiana%202020-5-14%20WEB%20VERSION.pdf.
[14] Murray Skees, "A New Traditional Theory: Fetishizing Big Data Analytics," *Constellations* 29, no. 2 (June 2022): 146–60, https://doi.org/10.1111/1467-8675.12541.

replace or limit Commerce and other organizations from seeking to understand lived experiences and communities' understandings of the problems they face. Commerce should explicitly investigate, acknowledge, address, and explain the limits of their datasets and generative AI. This is especially important as Commerce considers expanding its open data assets to be AI accessible, as, without proper precautions, improper outputs of such technology may be associated with the authority of government.

Additionally, a central function of generative AI is making predictions based on the data it was trained on, which creates the risk of assumptions and profiling individuals based on their identity. A study done by the European Union Panel for the Future of Science and Technology explains that "by correlating data about individuals to corresponding classifications and predictions, AI increases the potential for profiling, namely, for inferring information about individuals or groups, and adopting assessments and decisions on that basis".[15] This possibility raises issues of equity, as well as autonomy and freedom from discrimination which Commerce should be aware of when building the tools to further the progress of generative AI. This is especially important in certain high-risk domains like law enforcement, where the use of AI might pose too much of a risk to self-determination and autonomy.[16]

The EU's AI Act[17] takes steps mitigate the harms of AI in their risk based approach which deems the use of AI as high risk in scenarios such as "essential private and public services (e.g. credit scoring denying citizens opportunity to obtain a loan)", "educational or vocational training, that may determine the access to education and professional course of someone's life (e.g. scoring of exams)", etc. and in turn places certain obligations, on use of AI in these contexts, to limit their possible harms.[18] Commerce should look at current examples, nationally and internationally, that seek to recognize the kinds of data that pose a risk to autonomy, profiling, discrimination, and hallucinations in generative AI, or seem likely to lead to inequitable outcomes, and work to identify, note, and mitigate these harms.[19] While the stated goals of the Department of facilitating the ability for AI to easily access accurate data do not lead to all these harms, the Department should not only consider its intended outcomes for the data it opens but also all possible outcomes.

Another element that is important to consider when thinking about equitable outcomes is who will be able to make use of the open data Commerce is developing and publishing. The American Civil Liberties Union reports that along with data being "discriminatory or unrepresentative for people of color, women, or other marginalized groups […] the tech industry's lack of representation of people who understand and can work to address the potential harms of these technologies only exacerbates" the problem.[20] This indicates that while the focus on open data is fundamental for AI innovation, releasing data in AI accessible formats to be used to train generative AI is not enough to have truly open and equitable access to the benefits of data, especially when it lacks data that

---

[15] Giovanni Sartor and Francesca Lagioia, "The Impact of the General Data Protection Regulation on Artificial Intelligence.," 2020, https://data.europa.eu/doi/10.2861/293.
[16] Jenna Burrell and Jacob Metcalf, "Introduction for the Special Issue of 'Ideologies of AI and the Consolidation of Power': Naming Power," *First Monday*, April 14, 2024, https://doi.org/10.5210/fm.v29i4.13643.
[17] "AI Act | Shaping Europe's Digital Future," June 12, 2024, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.
[18] *Ibidem*.
[19] Moreover, AI can be exploited for malicious purposes such as deepfakes (particularly concerning when it involves elections), spreading misinformation, surveillance and manipulating public opinion. See, for example, how a government agency looks at it (https://www.gao.gov/products/gao-24-107292) and the role of AI in disinformation and misinformation in public health (https://www.bu.edu/ceid/2024/04/25/how-can-we-tackle-ai-fueled-misinformation-and-disinformation-in-public-health/)
[20] Olga Akselrod, "How Artificial Intelligence Can Deepen Racial and Economic Inequities | ACLU," *American Civil Liberties Union* (blog), July 13, 2021, https://www.aclu.org/news/privacy-technology/how-artificial-intelligence-can-deepen-racial-and-economic-inequities.

addresses concerns of marginalized groups. For instance, most current LLMs were developed from corpora skewed considerably towards the English language.[21]

Commerce should take note, and intentionally release data such that it is accessible to speakers of multiple languages and can facilitate models that are accurate in multiple languages. This is especially important for the United States as according to the Census, in 2019 "nearly 68 million people spoke a language other than English at home",[22] with 13% of the population speaking Spanish at home.[23] Without due consideration of the power, biases, and motivations of the AI industry, AI built on accessible open data will likely serve to benefit and further the influence of interests that are far from representative of the country's diverse demographics, and more broadly that of the public. In opening its data, Commerce should actively seek to counteract this by understanding and prioritizing the needs of those traditionally excluded from benefiting from data and AI.

Another topic that should be considered is consent. Since the U.S. still lacks proper privacy rights regulation at the federal level, people living in the country have different rights when it comes to data collection by government and companies. Commerce already has the data and should consider features that allow people to opt out of having their data used to train AI models whenever possible, investing in "consentful tech."[24] Global examples include projects such as My Data Rights in Africa,[25] Decode in Europe,[26] and the Indigenous Peoples' Rights in Data.[27] The Office of the Privacy Commissioner of Canada released principles[28] for more responsible generative AI technologies, stating that when relying on consent, it is crucial to ensure that consent is valid, informed and meaningful. Consent in this context still is a challenging process, so advancements in regulation are important to counter the ongoing issues of reaching meaningful consent.

Lastly, as technology policy is an ever-changing effort, Commerce must consult with and invite specialists, community members, advocates, scholars, as well as policymakers at the local level, to help frame its AI policies and data ethics frameworks. A multistakeholder approach to governing and regulating AI might be helpful to federal agencies (see more about it in section 4 – Partnerships). By sharing experiences and lessons learned between federal agencies and with civil society organizations, grassroots movements, communities, academia and companies, Commerce can build a robust, human-rights based strategy to release its open data assets.

2. Data dissemination standards
Responding to Q1, Q2 and Q4.

---

[21] SMC Spain. "Reactions: The Prime Minister Announces the Design of a Foundational Model of an Artificial Intelligence Language Trained in Spanish." SMC España, February 2024. https://sciencemediacentre.es/en/reactions-prime-minister-announces-design-foundational-model-artificial-intelligence-language.
[22]United States Census Bureau, *Nearly 68 Million People Spoke a Language Other Than English at Home in 2019*, December, 2022, https://www.census.gov/library/stories/2022/12/languages-we-speak-in-united-states.html.
[23] Sonia Thompson. "The U.S. Has the Second-Largest Population of Spanish Speakers-How to Equip Your Brand to Serve Them." Forbes, February 2024. https://www.forbes.com/sites/soniathompson/2021/05/27/the-us-has-the-second-largest-population-of-spanish-speakers-how-to-equip-your-brand-to-serve-them/.
[24] "What Is Consentful Tech?" The Consentful Tech Project, March 2022. https://www.consentfultech.io/.
[25] "Our Data, Our Rights." My Data Rights, December 2021. https://mydatarights.africa/.
[26] "Giving People Ownership of Their Personal Data." DECODE, January 2020. https://decodeproject.eu/.
[27] "Data Rights." Global Indigenous Data Alliance, 2023. https://www.gida-global.org/data-rights.
[28] Office of The Privacy Commissioner of Canada, P*rinciples for responsible, trustworthy and privacy-protective generative AI technologies*, https://www.priv.gc.ca/en/privacy-topics/technology/artificial-intelligence/gd_principles_ai/

Commerce should take note of the data dissemination frameworks being considered and implemented internationally. Simultaneously, awareness of how standards influence not just the structure of data-producing environments but also ensure compliance through funding criteria, partnerships, and recognition is important. A technical focus that is driven purely by standards might compromise authentic community engagement in processes that require conformity with limited choices.

In their report "Data Analytics and AI in Government Project Delivery" the United Kingdom identifies.[29] The FAIR principles,[30] widely covered by scholarship,[31] are also used by the EU to guide their assessment of the meta-data's quality associated with the data made available by various member countries.[32] It's data-centric approach can become a problem, since mere compliance with FAIR principles does not guarantee the absence of biases or societal injustices in open government data assets.

Additionally, Commerce can look at principles developed by the International Indigenous Data Sovereignty Interest Group, a "network of nation-state based Indigenous data sovereignty networks and individuals", called the CARE Principles (Collective Benefit, Authority to Control, Responsibility, Ethics).[33] These principles were developed to complement the FAIR principles which do not fully "engage with Indigenous Peoples' rights and interests."[34] Specifically, the CARE principles directly contend with how power and historical context shape who can benefit from open data -- and who is left behind. The CARE principles bring a people-and-purpose orientation to data governance, complementing the data-centric FAIR principles. They advocate for the "right to create value from Indigenous data in ways that are grounded in Indigenous worldviews and realise opportunities within the knowledge economy" focusing on "the crucial role of data in advancing Indigenous innovation and self-determination."[35]

Early adopters of the CARE principles (such as the Smithsonian Institution and the Open Data Charter), point out ways organizations can implement data dissemination guidelines more rooted in the public interest. Commerce can consider these principles as part of the process of providing open government data assets to the public, especially mechanisms to include voices rarely heard, create equitable outcomes, and improved data governance.

When releasing data for AI training models, Commerce should consider accessibility for people with disabilities both in the formats they release data in as well as how they report and create visualizations for the data considering "users with visual, auditory, motor, or cognitive

---

[29] Gov.UK, *Data Analytics and AI in Government Project Delivery,* March, 2024 https://www.gov.uk/government/publications/data-analytics-and-ai-in-government-project-delivery/data-analytics-and-ai-in-government-project-delivery

[30] It is worth noting that FAIR data is not synonymous with "open data". For example, data can be FAIR while still having restrictive access and reuse. FAIR principles are often criticized for its data-centric nature and lack of specific stakeholder focus. See, for example: Ugochukwu, Albert I., and Peter WB Phillips. "Open data ownership and sharing: Challenges and opportunities for application of FAIR principles and a checklist for data managers." *Journal of Agriculture and Food Research* (2024): 101157.

[31] Mark D. Wilkinson et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship," Scientific Data 3, no. 1 (March 15, 2016): 160018, https://doi.org/10.1038/sdata.2016.18.

[32] Metadata Quality," data.europa.eu, March 2024, https://data.europa.eu/mqa/methodology?locale=en#inline-nav-2

[33] Nico Riedel, Miriam Kip, and Evgeny Bobrov. "ODDPub – a Text-Mining Algorithm to Detect Data Sharing in Biomedical Publications." *Data Science Journal* 19, no. 1 (October 2020): 42. https://doi.org/10.5334/dsj-2020-042.

[34] Research Data Alliance International Indigenous Data Sovereignty Interest Group. (September 2019). "CARE Principles for Indigenous Data Governance." The Global Indigenous Data Alliance. GIDA-global.org.

[35] *Ibidem*.

disabilities."[36] Specifically, when creating human-readable data, Commerce should consider people with disabilities every step of the way, from how they make color and font choices to how visualizations are labeled and formatted. Even common ways of making data available may be inaccessible for people with disabilities. For instance, a common file type, PDFs, "inherently inaccessible to web users who are visually impaired or blind" because it's difficult to adjust colors and font sizes and they lack Alt Text descriptions.[37] The Center for Democracy and Technology reports that disability issues are rarely centered in tech policy spaces due to many factors explained in their report including "misconception of monolithic disability identity leading to misunderstanding of technology harms and remedies", "limited representation of individuals with disabilities", and "few formal and informal bridges between technology and disability advocates."[38] Commerce should use this opportunity to engage with disability advocacy groups and civil society organizations to understand and address the unique ways AI technology can benefit and harm people with disabilities.

It is also important that Commerce is "documenting the creation and use of data sets" in a standardized and consistent manner, which can be guided by a set of questions prominent AI researchers developed to be considered when creating a datasheet to document the creation of each data set.[39] Documentation for each dataset should cover all stages of the process including motivation for creating the data set, what is included in the data, how the data was collected, any processing or cleaning of the data, and what contexts the data should and should not be used for. For instance, any missing or synthetic data should be noted, as use of such data would influence models being trained. Documenting and explaining in a human understandable form what the data should not be used for is necessary to prevent unintentional misuse of the data and to create a means of accountability for harmful misuses of data. This documentation should not be created through automation as it also serves as a necessary process for human reflection on datasets that are being created and shared. Beyond traditional codebooks, this documentation should include information about the motives behind the creation of the dataset and any decisions what went into shaping its development.[40]

In terms of licenses and practices, a recent example worth noting and that can help Commerce design its policies is the "Recommended Best Practices for Better Sharing of Climate Data" by Creative Commons.[41] They explain that having "comprehensive, clear (plain language), and machine-readable metadata maximizes the reusability [...] by enabling replication and integration across different contexts. It makes climate data findable and improves search-engine optimization (SEO) for federated search engines, as well as for [...] organizational internal search engines. It boosts interoperability by providing qualified references to other (meta)data."[42] Other known

[36] Nancy Shin. Data visualizations for everybody: A lesson on accessibility, n.d. https://dataservices.library.jhu.edu/wp-content/uploads/sites/41/2024/03/24LDW_AccessibleDataViz_02-2024_NancyShin.pdf.
[37] "Usability and Accessibility Issues with Pdfs." Web Strategy, March 2023. https://webstrategy.med.wisc.edu/2022/08/17/usability-and-accessibility-issues-with-pdfs/.
[38] Henry Claypool et al., "Centering Disability in Technology Policy," *Center for Democracy and Technology*, December 2021, https://cdt.org/wp-content/uploads/2021/12/centering-disability-120821-1326-final.pdf.
[39] Timnit Gebru et al., "Datasheets for Datasets" (December 1, 2021), http://arxiv.org/abs/1803.09010.
[40] *Ibidem*.
[41] Taylor Campbell, "Recommended Best Practices for Better Sharing of Climate Data," Creative Commons, January 29, 2024, https://creativecommons.org/2024/01/29/recommended-best-practices-for-better-sharing-of-climate-data/.
[42] *Ibidem*.

examples are W3C DCAT specification,[43] ODI Open Data Rights Statement Vocabulary,[44] ISO standards, and the Global Climate Observing System (GCOS).[45]

Another recommendation to consider in this process is to have proper workforce development in the federal agency and enough resources to keep these open data platforms up to date, functional and usable. Establishing a tool or channel of communication, easily findable and easy to use, for people to suggest improvements, report problems and provide overall feedback is important, as it is to have processes of accountability in place to deal with harms and reported issues in a timely, ethical manner.

3. Data integrity and quality
Responding to Q2.

Making data available to be used in the training of AI models raises novel privacy concerns that need to be considered and addressed. First, Generative AI relies on the use of mass amounts of personal, sensitive information which can run counter to effective protection of privacy rights. The EU Panel for the Future of Science and Technology study explains that "there is indeed a tension between the traditional data protection principles – purpose limitation, data minimisation, the special treatment of 'sensitive data', the limitation on automated decisions and the full deployment of the power of AI and big data. The latter entails the collection of vast quantities of data concerning individuals and their social relations and processing such data for purposes that were not fully determined at the time of collection".[46] Commerce should be aware of this tension, and be sure to center the wellbeing of citizens and their interest in privacy and autonomy when deciding what data to publish and in what formats. This is particularly important as researchers have found that AI systems have the ability to"extract information from data, spot patterns and predict trends means that innocuous information can be mined to the point of relevance and intimacy".[47]

Commerce should consider how its technical privacy protection techniques, such as Differential Privacy, will be impacted by AI's ability to synthesize data across datasets, possibly identifying, or making accurate (or even inaccurate but believed) assumptions about information individuals would want to remain private. In addition, it is necessary to make it clear to individuals the intended uses of the collected data, as well as the conditions for sharing it with other agencies and third-party organizations.

Even if, on its face, Commerce's data does not seem to concern privacy, the uses that people and organizations will make of the data, and the models created based on them can have long-lasting impact, both positive and negative. For this reason, it is key to consider previous examples of privacy issues in data believed to be anonymized, and how those issues were (or were not) fixed. For over a decade, research has shown that it is possible to de-anonymize and identify individuals. In 2013, a study published in *Scientific Reports* analyzed anonymized mobile phone

[43] See https://www.w3.org/TR/vocab-dcat/.
[44] Leigh Dodds, "Open Data Rights Statement Vocabulary," July 2013, https://schema.theodi.org/odrs/.
[45] See https://council.science/member/global-climate-observing-system-gcos/.
[46] Giovanni Sartor and Francesca Lagioia, "The Impact of the General Data Protection Regulation on Artificial Intelligence.," 2020, https://data.europa.eu/doi/10.2861/293.
[47] Vidushi Marda, "Artificial Intelligence Policy in India: A Framework for Engaging the Limits of Data-Driven Decision-Making," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, no. 2133 (October 15, 2018): 20180087, https://doi.org/10.1098/rsta.2018.0087.

data.[48] The researchers showed that with just four spatiotemporal points (i.e., locations at specific times), they could uniquely identify 95 per cent of the individuals in a dataset of 1.5 million people. Other famous examples include studies that identified participants in the personal genome project matching names and contact information to publicly available profiles,[49] and the de-anonymization of medical data combining the anonymized data with publicly available information (e.g., ZIP code, birth date, and gender).[50] Federal agencies must adopt measures that prevent data de-anonymization related to respondents' personal identifiable information.

Without proper protections in place for private information, and explicit and enforced limitations on the release of that information such that it cannot be re-identified, it will be difficult to safeguard digital rights and facilitate efficient and truly open access to government data.[51]

4. Partnerships
Responding to Q1.

The U.S. has a thriving ecosystem of civil society organizations working on digital rights[52] and public interest technology, domestically and globally. The country also has well-established universities, research centers and think tanks dedicated to producing knowledge on AI and open data in various disciplines. Commerce should map relevant stakeholders and bring them into the process of establishing ethical open government data sharing. Public consultations are an excellent starting point that can and should be complemented by discussions with those who will use the data and those affected by the data made available to train AI models.

To bring people into the loop, it is important to make the process clear – countering AI black boxes[53] and their inherent opacity. While transparency must be one guiding principle of Commerce's plan, it is crucial to use transparency neither as an end in itself nor just a performative act[54] that would potentially diminish accountability efforts. Commerce should consider bringing in specialists in policy, civic engagement, data practices, explainability, and trust[55] to help steer this process. Specialists here can be understood as people with different degrees of education, practice and lived experience, not excluding community members and knowledge deemed non-academic.[56]

---

[48] Yves-Alexandre de Montjoye et al., "Unique in the Crowd: The Privacy Bounds of Human Mobility," Scientific Reports 3, no. 1 (March 25, 2013): 1376, https://doi.org/10.1038/srep01376.
[49] Latanya Sweeney, Akua Abu, and Julia Winn, "Identifying Participants in the Personal Genome Project by Name," SSRN Scholarly Paper (Rochester, NY, April 29, 2013), https://doi.org/10.2139/ssrn.2257732.
[50] Daniel Barth-Jones, "The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now," SSRN Scholarly Paper (Rochester, NY, July 1, 2012), https://doi.org/10.2139/ssrn.2076397.
[51] World Wide Web Foundation, "OPEN DATA BAROMETER REPORT – FROM PROMISE TO PROGRESS," 2018, https://opendatabarometer.org/doc/leadersEdition/ODB-leadersEdition-Report.pdf.
[52] See, for example, the organizations working on privacy and civil liberties compiled here: https://bja.ojp.gov/program/it/privacy-civil-liberties/agencies-org/pcl-organizations#0-0. Other examples include the Coalition for Independent Technology Research (https://independenttechresearch.org/) and Access Now and the conference Rightscon (https://www.accessnow.org/).
[53] Saurabh Bagchi, "What Is a Black Box? A Computer Scientist Explains What It Means When the Inner Workings of AIs Are Hidden," The Conversation, May 22, 2023, http://theconversation.com/what-is-a-black-box-a-computer-scientist-explains-what-it-means-when-the-inner-workings-of-ais-are-hidden-203888.
[54] Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989. https://doi.org/10.1177/1461444816676645.
[55] Dwork, C., & Minow, M. (2022). Distrust of Artificial Intelligence: Sources & Responses from Computer Science & Law. *Daedalus* 2022; 151 (2), p. 309–321. https://doi.org/10.1162/daed_a_01918
[56] See, for example: Brandusescu, A., & Reia, J. (Eds.). (2022). Artificial intelligence in the city: Building civic engagement and public trust. Centre for Interdisciplinary Research on Montreal, McGill University. https://doi.org/10.18130/9kar-xn17.

Government agencies must proactively create and release data for the benefit of all, particularly through engagement and collaboration with marginalized groups and communities.[57]

Another way to seek partners and comprehensive knowledge is to learn from and participate in international multistakeholder forums, organized or not by the United Nations (UN), dedicated to internet governance, open data, statistics and AI. These venues can offer years of accumulated knowledge, helping Commerce find stakeholders to advise in the current process and learn about recent trends from people who have been studying open data practices and emerging AI systems for years.[58] Commerce can consider its own workshops and permanent spaces of discussions, in which people representing organizations from different sectors come together to assess what has been done, propose changes and work towards future initiatives.

Commerce must also keep in mind the limitations of such international forums where participation between stakeholder groups often is skewed and actively work to bring a nuanced and balanced range of diverse expertise.[59] Thus, it is important to establish "a regular communication channel with civil society and non-profit organizations through workshops, public hearings, public consultations, studies, grants, and online engagement" and "include underrepresented communities to ensure AI solutions are inclusive and to avoiding potential harm (such as Indigenous peoples, Black people, people of colour, and members of the LGBTQIA+ community)."[60]

Before engaging in AI-related partnerships, Commerce should consider the threshold and standards for companies and organizations it partners with. From the legal and administrative procedures to establish partnerships to the diversity of stakeholders and individuals, a set of guidelines would be helpful. For example, not partnering with companies that have a record of human rights violations[61] (in the U.S. and abroad) and partnering with people and organizations that are rarely heard in the open data and AI spaces. When deciding who to collaborate with in industry, commerce should consider their actions involving "the gathering, use and commercialization of personal data; freedom of expression; facilitating the spread of hate speech, misinformation, political extremism, terrorism, electoral manipulation and the suppression of democratic dissent; the impacts of content moderation and encryption; discrimination and other human rights abuses resulting from algorithmic bias; and impacts on at-risk groups including children and human rights defenders".[62]

Finally, tools, mechanisms and principles to assess the risks of AI models, which can also guide Commerce's partnerships, can be relevant guides – especially when other federal agencies intend to use the data made available by Commerce in automated decision-making.[63] Furthermore, clearly specify certain conditions under which the data can and cannot be used (such as high-risk automated processes). Internationally, examples of risk and impact assessments include the Algorithmic Impact Assessment (AIA) in Canada, a mandatory risk assessment tool intended to

---

[57] World Wide Web Foundation, "OPEN DATA BAROMETER REPORT – FROM PROMISE TO PROGRESS," 2018, https://opendatabarometer.org/doc/leadersEdition/ODB-leadersEdition-Report.pdf.
[58] Other than the aforementioned Rightscon, two other relevant forums are the UN Internet Governance Forum – IGF (https://www.intgovforum.org/en) and the UN World Data Forum (https://unstats.un.org/unsd/undataforum/).
[59] Sambuli (2021): https://www.carnegiecouncil.org/media/article/five-challenges-with-multistakeholder-initiatives-on-ai.
[60] Novartis Foundation,"AI4 Healthy Cities" (January 2022), https://www.datocms-assets.com/57996/1642518870-ai4healthycities.pdf.
[61] USA: Failing to do right: The urgent need for Palantir to respect human rights. https://www.amnesty.org/en/documents/amr51/3124/2020/en/
[62] Claire O'Brien, Rikke Jørgensen, and Benn Hogan, "Tech Giants: Human Rights Risks and Frameworks," SSRN Scholarly Paper (December 15, 2020), https://doi.org/10.2139/ssrn.3768813.
[63] U.S. Chief Information Officers Council, *Algorithmic Impact Assessment*, https://www.cio.gov/aia-eia-js/#/

determine the impact of public sector use of automated decision-making systems;[64] the UK government's Artificial Intelligence Impact Assessment (AIIA);[65] and UNESCO's Ethical Impact Assessment.[66]

Overall, Commerce's partnerships should be multistakeholder, inclusive, human-rights-based and close to communities to understand how to best serve people while innovating and being beneficial to other government agencies.

5. Questions not being asked: Climate action and environmental justice

Commerce should also be more concerned with the environmental impact of the data they collect and share, as well as the environmental impact of releasing that data to train AI models. The environment, and questions of sustainability that arise from mass amounts of data collection, storage, and use, should all be central to the steps taken to move forward with creating, storing, and sharing new forms of this data, particularly those that are made for training AI models.

Since the social implications of AI are transnational, global examples of best practices and regulatory pathways are useful. In their Guide on the Use of Generative Artificial Intelligence,[67] the Government of Canada identifies environmental harms that may arise explaining that "the development and use of generative AI systems can be a significant source of greenhouse gas (GHG) emissions and water usage. These emissions come not only from the computer used to train and operate generative models but also from the production and transportation of servers that support AI programs. In addition, data centres are energy-intensive and consume vast quantities of water for on-site cooling and off-site electricity generation".[68]

Acknowledging the place of big data and AI in the current climate crisis, reflected on current bills in the U.S. and abroad, is the first step to a responsible, environmentally just approach. Commerce should compile and disclose the footprint of its data initiatives and AI partnerships, as well as study the demands for large data centers from the perspective of residents being directly affected by such infrastructures. An example is the Virginia Data Center Reform Coalition, "urging state lawmakers to study the cumulative effects of data center development on Virginia's electrical grid, water resources, air quality, and land conservation efforts, and to institute several common-sense regulatory and rate-making reforms for the industry".[69]

Commerce should be especially aware of the environmental impact of its data on marginalized groups as researchers have found that "increasing the environmental and financial costs of these [AI] models doubly punishes marginalized communities that are least likely to benefit from the progress achieved by large LMs and most likely to be harmed by negative environmental

---

[64] Created by the Treasury Board of Canada Secretariat, the AIA tool is a questionnaire that determines the impact level of an automated decision-system. It is composed of 51 risk and 34 mitigation questions: https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

[65] Centre for Data Ethics and Innovation and Department for Science, Innovation and Technology, *RAI Institute: Artificial Intelligence Impact Assessment (AIIA),* April 2024, https://www.gov.uk/ai-assurance-techniques/rai-institute-artificial-intelligence-impact-assessment-aiia

[66] UNESCO Global AI Ethics and Governance Observatory, *Ethical Impact Assessment,* 2023, https://www.unesco.org/ethics-ai/en/eia.

[67] Government of Canada, *Guide on the use of Generative Artificial Intelligence*, June, 2024, https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/guide-use-generative-ai.html.

[68] Ibidem.

[69] Piedmont Environmental Council, "Virginia Data Center Reform Coalition," January 2022, https://www.pecva.org/work/energy-work/data-centers/virginia-data-center-reform-coalition/.

consequences of its resource consumption. At the scale we are discussing, the first consideration should be the environmental cost".[70]

Moreover, Commerce mobilizing its resources to indiscriminately release its data in AI accessible formats would play a role in the intensive resource needs of big data. Federal agencies should keep in mind how their data practices and AI adoption will contribute to water shortages, displacement, and the overextraction of minerals as it would make it extremely easy for Generative AI models to be trained on larger and larger datasets.[71] Even if, in the process, releasing this data leads to better understandings of environmental issues and how to address them, these means run directly counter to this goal. Instead of mobilizing time and resources to making as much governmental data as possible AI accessible, Commerce should be considerate and selective and focus on high impact data and data needed by communities, with a particular focus on data that can be used to understand social, economic, environmental, and other pressing problems.

More and more data sets and larger data centers are not necessarily better. Instead, to have the most positive impact, Commerce should use its influence, expertise, and resources to prioritize technology and practices that have sustainability and climate action at its center.

Respectfully submitted,

Jess Reia, PhD
Assistant Professor of Data Science
Faculty Lead, Karsh Institute's Digital Technology for Democracy Lab
University of Virginia
1919 Ivy Road, office 336
reia@virginia.edu

Rachel Leach
Research Assistant
Data Justice and Climate Resilience in the Global Automotive Industry Project
University of Virginia

[70] Emily M. Bender et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event Canada: ACM, 2021), 610–23, https://doi.org/10.1145/3442188.3445922.
[71] Mél Hogan and Théo Richer, "Extractive AI," Centre for Media, Technology and Democracy, 2024, https://www.mediatechdemocracy.com/climatetechhoganlepagericher.