CENTER FOR DEMOCRACY & TECHNOLOGY

February 9, 2021

Advisory Committee on Data for Evidence Building
Via regulations.gov

Re: Request for Comments for the Advisory Committee on Data for Evidence Building

The Center for Democracy & Technology (CDT) submits these comments to highlight steps the Advisory Committee on Data for Evidence Building should take to help bolster the equitable provision of government services, community trust in federal data, and individual and organizational privacy. Coordinated, data-driven action across interdependent agencies is essential to providing vital services. Data can help reveal inequitable access to services and data, support steps to increase economic mobility, and raise under-explored questions about the role of race and racism in the design and implementation of government programs and policies. CDT believes that comments led by the Annie E. Casey Foundation in this Docket succinctly summarize the value of the Advisory Committee's work.

However, data use also poses risks to individual and organizational privacy and autonomy, and CDT urges the Advisory Committee to commit to principles of responsible data governance, stakeholder engagement, equity, and transparency in federal agencies' collection and use of data.

Responsible Data Governance
In order to promote equity and protect privacy, the Advisory Committee should ensure that federal evidence building is supported by responsible data governance. Data governance is "the overall management of data, including its availability, usability, integrity, quality, and security,"[1] and includes people, processes, and structures that are responsible for data and technology. The Advisory Committee should ensure that federal evidence building incorporates key ethical data practices such as the recommendations of the Commission on Evidence-Based Policymaking.[2] Such practices should include:

- *Data minimization:* Agencies should collect, use, retain, and share only the data required to fulfill a clear and specific purpose, so as to minimize the risks from unauthorized access or use of data out of context. Additionally, secure and appropriately limited data sharing and user access can assist in data minimization by

[1] Corey Chatis and Kathy Gosa, Communicating the Value of Data Governance, SLDS Issue Brief (2017).
[2] Commission on Evidence-Based Policymaking, The Promise of Evidence-Based Policymaking (2017).

ensuring that the same data is not collected and stored multiple times for different purposes.[3]

- *Data ownership:* Defining who has the ultimate control, responsibility, and legal rights over the data is an important decision that is best made early and documented in formal agreements between the agency, the data source, and recipients of the data.[4]
- *Data retention, storage, and deletion:* Agencies should establish transparent rules and processes to ensure secure storage and retention periods that are no longer than necessary for the purpose for which the data is processed to minimize the risks that come with amassing unnecessary data.[5]
- *Data sharing:* Agencies will need to consider whether sharing is appropriate, necessary, and consistent with users' expectations and will need to develop clear policies that govern the roles, responsibilities, and processes for sharing. This includes requirements around access, privacy, storage, use, and deletion.[6]
- *User access:* Limiting user access to only individuals who have a clear need for the data can help agencies ensure privacy protection and minimize the likelihood of inappropriate data access or misuse.[7] As the Commission on Evidence-Based Policymaking recommended, the Advisory Commission should consider adopting a "tiered-access system" to limit user access to data based on sensitivity,[8] requiring risk assessments for public disclosures,[9] and establishing disclosure review boards.[10]
- *Data quality:* Agencies have an ethical obligation to ensure the accuracy of the data they use. Otherwise, any insights gleaned from that data or actions taken based on that data may be misguided and do more harm than good. Agencies should consider adopting mechanisms for users to view and request the correction and deletion of information held about them.
- *Documentation:* To provide transparency and accountability, agencies should document their policies and procedures for data collection, data use, data sharing with vendors and other third parties, and decision-making based on the data.[11]

---

[3] Joanna Grama, Protecting Privacy and Information Security in a Federal Postsecondary Student Data System (2019).
[4] Brian Bollier, The Promise and Peril of Big Data (2010).
[5] White House Big Data and Privacy Working Group, Big Data: Seizing Opportunities, Preserving Values (2015).
[6] John Fantuzzo et al., The Integrated Data System Approach: A Vehicle to More Effective and Efficient Data-Driven Solutions in Government (2017).
[7] Omer Tene and Jules Polonetsky, Big Data for All: Privacy and User Control in the Age of Analytics, 11 Northwestern Journal of Technology and Intellectual Property 240–72 (Apr. 2013).
[8] Commission on Evidence-Based Policymaking, *supra* note 2, at 41.
[9] *Ibid.* at 61.
[10] *Ibid.* at 50.
[11] Joel Reidenberg et al., Privacy and Cloud Computing in Public Schools, FLASH: The Fordham Law Archive of Scholarship and History Book 2 (2013).

The Commission on Evidence-Based Policymaking also envisioned establishing the National Secure Data Service (NSDS) for temporarily linking existing datasets, which would be governed by a Steering Committee composed of diverse stakeholders with an established process for assessing requests for linking datasets.[12] The Commission also recommended that disclosures from federal data be subject to "strict data minimization techniques to ensure researchers accessing combined data will use datasets with as much information removed as is possible while still meeting the research need. When two or more datasets will be combined, only a narrow group of qualified and trained employees will have access to direct identifiers to conduct the linkage."[13] Regardless of whether the NSDS is eventually established, the Advisory Committee should ensure that there are clear structures for responsible data governance.

Stakeholder Engagement

As noted in the comments led by the Annie E. Casey Foundation, it is essential that the Advisory Committee engage stakeholders whose data is being collected and who utilize the services supported by that data. Data and technology initiatives benefit from diverse perspectives, surfacing potential problems and developing frameworks that work for a broad cross-section of users. Stakeholder engagement will also increase buy-in and trust in how data and technology are used, which can increase faith in federal data more broadly. Moreover, agencies are more likely to encounter pushback on how data is being used if they do not engage stakeholders. In the event of a breach or other issue, stakeholders are more likely to be understanding if they had buy-in at the outset, seeing firsthand that meaningful steps were taken to put protections in place.[14] Stakeholder engagement can range from informational to advisory, or even to giving stakeholders decision-making authority, depending on the topic and capacities involved.

Equity

Data and technology provide potential benefits to individuals as well as the public good. However, these benefits will only be realized if the collection, analysis, and use of data are designed intentionally to meet these goals and minimize potential bias. To this end, it is important that the Advisory Committee identify and address the ways in which data and technology use could inadvertently create, entrench, or worsen inequities or have other unintended consequences.

Certain data practices may create inequities in policymaking if the data elements are biased, and including those data elements in analyses may bias the outcomes towards (or against) particular groups. For example, in education, students of color are disciplined at a greater rate than their peers (both in terms of number of infractions as well as the severity of

[12] *Ibid.* at 81-84.
[13] *Ibid.* at 40.
[14] Ben Green and Lily Hu, The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning, in Machine Learning: The Debates Workshop at the 35th International Conference on Machine Learning (2018).

consequences), so using discipline data in certain analyses could result in the over- or under-identification of students of color, which could negatively affect their outcomes. Alternatively, using data from a non-representative sample and then applying the findings to the broader population can result in practices or policies that are not beneficial for certain populations within the broader community.[15]

The Advisory Committee should also consider other equity issues that can arise from the use of data and technology, especially the way an agency's authority to grant or deny benefits may influence an individual's willingness to exercise their data rights. If an agency that has the authority to grant or deny benefits is the same agency that controls an individual's data, the individual may lack the power or comfort to request access to, correct, or delete their information or to push back if their requests are not honored.[16]

Lastly, emerging technologies have the potential to exacerbate bias. For example, predictive analytics, particularly when machine learning is utilized, can significantly increase inequitable outcomes if bias is not accounted for in their design and evaluation.[17] For example, as noted above, students of color are disproportionately disciplined at a greater rate than their peers, so early warning systems that use discipline data to predict whether a student is on track or at risk of dropping out of school will identify more students of color.

Transparency & Secondary Data Uses

The Advisory Committee should seek to ensure transparency at all stages of the data lifecycle, from collection through analysis and use, to support data quality, create trust, and establish buy-in. Transparency is a broad concept but should cover, at minimum, data collection, use, storage, and decision-making. Specifically regarding decision-making, transparency includes visibility into how decisions are being made based on data, including methodology, decision-making processes, and the underlying data itself.

Transparency is a particular concern in evidence-based policymaking, when data may be re-used for additional purposes beyond the original intended use, potentially diverging from the scope of what the data subject was notified of or consented to. Secondary data use can become an issue with any data that is collected, including:

- Data that was collected for informational purposes and then is used for decision-making;
- Data that was originally not going to be shared with outside agencies, but then is shared externally;
- Data that was collected to support the individual but then is used for a collective purpose such as research; or

[15] Andrea Alarcon et al., Data & Civil Rights (2014).
[16] Randy Bean, A Rising Crescendo Demands Data Ethics and Data Responsibility, Forbes (Oct. 29, 2018).
[17] Andrew Cormack, A Data Protection Framework for Learning Analytics, Journal of Learning Analytics (2016).

- Data that was collected, aggregated, and used for systemic decisions but is then disaggregated and used to make decisions about individuals.

Secondary data use is especially pertinent to research and open data. Data is often collected across fields to track and support individual outcomes (e.g., test scores, health screenings), but may also be helpful for research to support the broader sector. Often, these research projects have not yet been identified at the time of the data collection, so consent can be difficult or impossible to collect. In some cases, de-identified or aggregate data could be used for research purposes and may pose less of a privacy risk, but de-identification must be done carefully by someone with proper training to minimize the risk that the data is re-identified, thus exposing the individuals to privacy loss, financial risk, or other harms.

Secondary data uses may have a particularly pronounced impact on underserved or marginalized communities. For example, in Pasco County, Florida, children's school records were shared with law enforcement without parental consent to create a "predictive policing" system.[18] That system incorporated school data, including discipline records, to identify "students who are at-risk of developing into prolific offenders."[19] As in many school districts, Black students and students with disabilities in Pasco County are twice as likely to be disciplined, which may increase their exposure to law enforcement due to the district's data sharing.[20]

The Evidence Act[21] already has some limitations on secondary uses, ensuring that "[d]ata information acquired by an agency under a pledge of confidentiality and for exclusively statistical purposes shall be used by officers, employees, or agents of the agency exclusively for statistical purposes and protected in accordance with such pledge."[22] The Act similarly prohibits "nonstatistical uses" of such data, including for "any administrative, regulatory, law enforcement, adjudicatory, or other purpose that affects the rights, privileges, or benefits of a particular identifiable respondent."[23] The Advisory Committee should ensure that federal evidence building adheres to that functional distinction. It should likewise provide guidance on responsible data governance, stakeholder engagement, and equity to guide agencies in determining whether they have a legal and ethical basis for secondary uses of data.

---

[18] Neil Bedi & Kathleen McGrory, Pasco's Sheriff Uses Grades and Abuse Histories to Label Schoolchildren Potential Criminals, Tampa Bay Times (Nov. 19, 2020), https://projects.tampabay.com/projects/2020/investigations/police-pasco-sheriff-targeted/school-data/.

[19] *Ibid.*

[20] *Ibid.*; *see also* F. Chris Curran, 'Early warning' Systems in Schools Can Be Dangerous in the Hands of Law Enforcement, The Conversation, https://theconversation.com/early-warning-systems-in-schools-can-be-dangerous-in-the-hands-of-law-enforcement-152701.

[21] Foundations for Evidence-Based Policymaking Act of 2018, Pub. L. No. 115-436, 132 Stat. 5529 (2019).

[22] 44 U.S.C. § 3572(b).

[23] 44 U.S.C. § 3572(d); 44 U.S.C. § 3561(8).

CENTER FOR DEMOCRACY & TECHNOLOGY

CDT applauds the efforts of the Advisory Committee and agencies across the federal government to use data ethically and equitably. We believe that evidence-based policymaking can be used to create more equitable government services while protecting individual and organizational privacy. We look forward to working with the Advisory Committee as it further considers these issues.

Sincerely,

Elizabeth Laird
Director, Equity in Civic Technology, CDT

Cody Venzke
Policy Counsel, Equity in Civic Technology, CDT