



Before the National Institutes of Health, Department of Health and Human Services

Comments of Box Commons on the Proposed Reorganization of the Division of Program Coordination, Planning, and Strategic Initiatives

Re: Notice of Proposed Reorganization, FR Doc. 2026-20891 (91 Fed. Reg. 65229, Oct. 14, 2026): establishing the Office of the NIH Chief Artificial Intelligence Officer (OCAIO) within DPCPSI

Submitted to: DPCPSIreorgcomments@nih.gov

Date: October 23, 2026

I. Summary

Box Commons supports establishing the OCAIO within DPCPSI. An NIH-wide priority deserves clear leadership, and placing the office in DPCPSI puts it next to the strongest model NIH has for AI-ready data: the Common Fund's Bridge to Artificial Intelligence (Bridge2AI) program.¹

Our one request: make the consent and provenance practices that Bridge2AI's projects already use the NIH-wide default for data used to develop AI. NIH would not need to invent anything: each recommendation names a practice an NIH-funded team already runs.

What the OCAIO can make standard	Who already does it
1. A withdrawal statement that covers future releases, downloaded copies, and trained models	Bridge2AI-Voice drops withdrawn participants from new releases. No project yet addresses trained models.
2. Tiered release by identifiability, with derived features first for biometric data	Bridge2AI-Voice (raw audio held back; derived features released under a data use agreement)
3. One machine-readable datasheet that records provenance and consent scope	Bridge2AI Standards Working Group;
4. Consent forms that name AI as a purpose, published openly	Bridge2AI-Voice; AI-READI
5. A written NIH definition of "ethically sourced"	Bridge2AI-Voice
	Bridge2AI ethics researchers have drafted one

II. Interest of the Commenter

Box Commons is a Wyoming nonprofit organized as a Section 501(c)(6) standards body. It focuses on clean audio data: consent, provenance, chain of custody, and accountability when people use

¹NIH Common Fund, *Bridge to Artificial Intelligence (Bridge2AI)*, <https://commonfund.nih.gov/bridge2ai>. The Common Fund is managed by the Office of Strategic Coordination within DPCPSI.

recorded voice to build AI. Our standards are technology-neutral and outcome-based. Voice is the hardest case for research data governance, because a recording can both identify the speaker and reveal their health. Bridge2AI-Voice, led from the University of South Florida, is building a shared dataset to test whether AI can detect conditions such as Parkinson’s disease, dementia, and mood disorders from the voice. Because voice is identifying, the project had to solve consent and release problems that every AI dataset will face. We have raised related themes with CDC, the Copyright Office, the FCC, and the Colorado Department of Law.²

III. The office already has this job

The recommendations fit inside duties NIH and the federal government have already assigned.

- **Federal Chief AI Officer duties.** OMB Memorandum M-25-21 directs each agency Chief AI Officer to “ensure that custom-developed AI code and the data used to develop and test AI are appropriately inventoried, shared, and released.” For AI that agencies buy, it encourages “documenting provenance of the data used to train, fine-tune, or operate the AI.”³
- **NIH data sharing policy.** The Data Management and Sharing Policy encourages investigators to address data sharing “in the informed consent process,” and says any “limitations on subsequent use of data” should be communicated to those who preserve and share it.⁴
- **NIH’s open question on AI.** NIH already considers “generative AI models, including model parameters,” trained on controlled-access data to be Data Derivatives, and has said it “intends to provide future guidance” on them.⁵

What is missing is a single place where consent scope and provenance get recorded for AI use. The updated Data Management and Sharing Plan format, effective for due dates on or after May 25, 2026, asks applicants to justify limits on sharing and to describe privacy protections. It does not ask what participants agreed to about AI.⁶ The OCAIO is well placed to close that gap.

IV. Recommendations

1. A withdrawal statement that covers trained models

What Bridge2AI does. The Bridge2AI-Voice consent form tells participants that once voice collection is complete, their “voice recording, and other research data cannot be removed from the database.”⁷ In practice, the project goes further than the form promises. Version 4.0.0 of its dataset (October 19, 2026) notes that “19 v3.1 participants are not in this version, because of changes to

²Comments of Box Commons, *Proposed Rules for the Automated Decision-Making Technology Act (SB 26-189) and the Chatbot Safety Act (HB 26-1263)*, 4 CCR 904-6, Colorado Department of Law (submitted Sept. 23, 2026).

³Office of Management and Budget, Memorandum M-25-21, *Accelerating Federal Use of AI through Innovation, Governance, and Public Trust* (Apr. 3, 2025), §§ 3(a)(i)(H), 2(d)(ii)(B), <https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-21-Accelerating-Federal-Use-of-AI-through-Innovation-Governance-and-Public-Trust.pdf>.

⁴NIH, *Final NIH Policy for Data Management and Sharing*, NOT-OD-21-013, <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html>.

⁵NIH, *Protecting Human Genomic Data when Developing Generative Artificial Intelligence Tools and Applications*, NOT-OD-25-081 (Mar. 28, 2025), <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-25-081.html>.

⁶NIH, NOT-OD-26-046, <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-26-046.html>, and *Data Management and Sharing Plan Format Page*, <https://grants.nih.gov/grants-process/write-application/forms-directory/data-management-and-sharing-plan-format-page>.

⁷University of South Florida, *Bridge2AI eConsent, English, Version 4* (effective Aug. 31, 2026), Bridge2AI-Voice REDCap repository, <https://github.com/eipm/bridge2ai-redcap/tree/main/docs/Adults/PDFs/en-us>.

their eligibility, consent or data-sharing status.”⁸ Withdrawn participants leave future releases.

The open question. No Bridge2AI project we reviewed states what happens to copies researchers have already downloaded, or to models already trained on a participant’s data. This is the question participants are most likely to ask, and NIH has already done half the work by treating trained models as Data Derivatives.⁹

What the OCAIO can make standard. Every NIH-supported dataset released for AI work should carry a short withdrawal statement covering three things: (a) future releases, (b) copies already distributed under data use agreements, and (c) models already trained. NIH need not pick one answer for (c); each dataset just has to state its answer plainly. The guidance NIH promised in NOT-OD-25-081 is a natural home for it.

2. Tiered release by identifiability, with derived features first for biometric data

What Bridge2AI does. Bridge2AI-Voice never released raw audio openly. Its first release stated that “audio waveforms were omitted, and only spectrograph data and other derived features are made available,” and that free-speech transcripts were removed.¹⁰ The current release, with 48,226 recordings from 1,370 participants, goes further. It withholds “highly identifying features such as speaker embeddings,” down-samples spectrograms “to make speaker identification more difficult,” and is available only to credentialed users who sign a data use agreement.¹¹ Raw audio is available separately; access “requires institutional sign off” and goes through a data access committee. The consent form sets up these tiers in advance: low-risk data may be open, while data with “heightened risk of re-identification,” including “certain voice data recordings,” goes through controlled access.¹²

AI-READI, another Bridge2AI project, adds license terms suited to AI. Users may train AI models, but may not “attempt to identify study participants,” and “the data may not be used for general LLM training purposes nor uploaded to a third party LLM vendor” (LLM meaning large language model).¹³

What the OCAIO can make standard. For biometric signals (voice, face, gait, retinal images), the NIH default should be: release derived features first, keep the most identifying representations back, and route raw signals through a data access committee with a published list of approved uses. Each tier should cite the consent language that permits it.

Why audio needs this. We have made the same point in other settings. Provenance attached when a recording enters the supply chain prevents problems that detection after the fact can only chase.¹⁴ Health data exchange has independent certification for cybersecurity, but nothing

⁸Bensoussan Y, et al., *Bridge2AI-Voice: An ethically-sourced, diverse voice dataset linked to health information*, v4.0.0 (PhysioNet, Oct. 19, 2026), <https://doi.org/10.13026/hh2a-kn84>.

⁹NIH, *Protecting Human Genomic Data when Developing Generative Artificial Intelligence Tools and Applications*, NOT-OD-25-081 (Mar. 28, 2025), <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-25-081.html>.

¹⁰Bridge2AI-Voice, v1.0 (Health Data Nexus, Nov. 27, 2024), <https://healthdatanexus.ai/content/b2ai-voice/1.0/>.

¹¹Bensoussan Y, et al., *Bridge2AI-Voice: An ethically-sourced, diverse voice dataset linked to health information*, v4.0.0 (PhysioNet, Oct. 19, 2026), <https://doi.org/10.13026/hh2a-kn84>.

¹²University of South Florida, *Bridge2AI eConsent, English, Version 4* (effective Aug. 31, 2026), Bridge2AI-Voice REDCap repository, <https://github.com/eipm/bridge2ai-redcap/tree/main/docs/Adults/PDFs/en-us>.

¹³AI-READI, *License* (documentation), <https://docs.aireadi.org/docs/3/preliminary/license>; Contreras J, et al., *License terms for reusing the AI-READI dataset*, v2.0 (Nov. 7, 2025), <https://zenodo.org/records/17555036>.

¹⁴Comments of Box Commons, *Music Streaming Fraud*, Notice of Inquiry, U.S. Copyright Office Docket No. 2026-6 (submitted Oct. 13, 2026).

comparable for data quality and provenance.¹⁵ Voice adds a third reason: the signal itself can identify a person, so data stewards have to decide release per representation, not per dataset.¹⁶

3. One machine-readable datasheet that records provenance and consent scope

What Bridge2AI does. The Bridge2AI Standards Working Group published criteria for AI-ready biomedical data across seven dimensions, including **Provenance** and **Ethics**. The criteria call for identifying “people and organizations responsible for obtaining and processing the data” and for ensuring that “the intended applications align with consent and institutional agreements.” They also caution against the CC0 public-domain dedication because “it fails to preserve provenance.”¹⁷ Bridge2AI maintains an open, machine-readable datasheet schema built on the “Datasheets for Datasets” model.¹⁸ Bridge2AI-Voice publishes, for every data file, a data dictionary listing “extraction parameters and software version,” along with recording metadata such as “audio sample rate and channel count, and device and browser information.”¹⁹

The burden today. The AI-READI team reports that its own datasheet “required involvement from over 15 persons and took over 2 months to finalize.” It concludes that “clear guidelines by funders, data repositories, and scientific journals will be needed.”²⁰

What the OCAIO can make standard. Adopt one NIH datasheet schema, built from Bridge2AI’s, for any dataset offered for AI development. It should include a short consent block covering whether AI development is a permitted use, which access tier applies, and the withdrawal statement from Recommendation 1. For audio and other sensor data, it should include the recording conditions that already appear in Bridge2AI-Voice’s metadata. Linking the datasheet to the Data Management and Sharing Plan lets investigators fill it in once, at collection, rather than rebuilding it at release.

4. Consent forms that name AI as a purpose, published openly

What Bridge2AI does. The Bridge2AI-Voice consent form states its purpose plainly: to learn whether voice biomarkers “could be screened and differentiated through AI and machine-learning algorithms.”²¹ The project publishes its consent forms from each site, including revised versions, in a public repository.²²

What the OCAIO can make standard. Add model language naming AI development as a purpose to NIH’s existing sample consent language for secondary research.²³ Encourage NIH-

¹⁵Comments of Box Commons, *Data Intermediaries and Approaches to Strengthen Public Health Data Exchange*, Docket No. CDC-2026-1519, 91 Fed. Reg. 59777 (submitted Sept. 30, 2026).

¹⁶See Comments of Box Commons, *Improving the Effectiveness of the Robocall Mitigation Database*, WC Docket Nos. 24-213 and 17-97, CG Docket No. 17-59 (filed Sept. 24, 2026), on authenticating voice content, not just the channel.

¹⁷Clark T, Caufield H, Parker JA, et al., *AI-readiness Criteria for Biomedical Data*, bioRxiv preprint, <https://doi.org/10.1101/2024.10.23.619844>.

¹⁸Bridge2AI, *Datasheets for Datasets, as LinkML Schema*, <https://github.com/bridge2ai/data-sheets-schema>.

¹⁹Bensoussan Y, et al., *Bridge2AI-Voice: An ethically-sourced, diverse voice dataset linked to health information*, v4.0.0 (PhysioNet, Oct. 19, 2026), <https://doi.org/10.13026/hh2a-kn84>.

²⁰Heinke A, et al., *Dataset documentation for responsible AI: analysis of suitability and usage for health datasets*, npj Digital Medicine 9:547 (2026), <https://doi.org/10.1038/s41746-026-02714-2>.

²¹University of South Florida, *Bridge2AI eConsent, English, Version 4* (effective Aug. 31, 2026), Bridge2AI-Voice REDCap repository, <https://github.com/eipm/bridge2ai-redcap/tree/main/docs/Adults/PDFs/en-us>.

²²Bridge2AI-Voice consent forms (USF, MIT, Vanderbilt University Medical Center, and Weill Cornell Medicine), <https://github.com/eipm/bridge2ai-redcap/tree/main/docs/Adults/PDFs>.

²³NIH Office of Science Policy, *Informed Consent for Secondary Research with Data and Biospecimens: Points to Consider and Sample Language for Future Use and/or Sharing*, <https://osp.od.nih.gov/wp-content/uploads/Informed->

supported projects that generate AI data to publish their consent forms alongside the dataset, so data users can check that their intended use falls within what participants agreed to.

5. A written NIH definition of “ethically sourced”

What Bridge2AI’s researchers found. Bridge2AI describes its goal as AI-ready data that is “ethically sourced.”²⁴ Researchers on the Bridge2AI-Voice project observed that ethically sourced data “is a concept fundamental to the Bridge2AI program,” yet “it has not been defined by the NIH.”²⁵ A later Bridge2AI ethics paper notes that the term “was not defined at the program’s onset” and offers a working account: “ethical sourcing includes documenting data provenance, articulating expectations for experts, and practices for ensuring data privacy, equity, and public benefit.”²⁶ Related work from the same team finds that “encryption, anonymization, and de-identification” have “proven to be more challenging with voice and speech data than with other biometric information.”²⁷

What the OCAIO can make standard. Adopt a short written definition of “ethically sourced,” starting from the work these researchers have done, and tie it to Recommendations 1 through 4 so the term means something NIH and its grantees can show in practice.

V. Conclusion

The OCAIO will shape how NIH and its grantees build AI for years. NIH’s own Bridge2AI teams have already worked out practical answers to the hardest questions, especially for voice. Making those answers the NIH default, recorded at the point of collection, would give the new office a strong start. Box Commons appreciates the opportunity to comment and would welcome the chance to help.

Respectfully submitted,

Brice Love

Founding Catalyst

Box Commons

c/o Northwest Registered Agent LLC, 30 N Gould St Ste N, Sheridan, WY 82801

info@boxcommons.org · boxcommons.org

Methodology and AI disclosure: Box Commons prepared this comment with an AI-assisted research and drafting workflow (Claude, Anthropic) covering the sources cited. Brice Love, Founding Catalyst, directed and approved the analysis, and we checked every quotation against its primary source. Box Commons takes full responsibility for this submission. Under Executive Order 14434, NIH may also say “Super Intelligence (SI).”

Consent-Resource-for-Secondary-Research-with-Data-and-Biospecimens.pdf.

²⁴NIH Common Fund, *Bridge to Artificial Intelligence (Bridge2AI)*, <https://commonfund.nih.gov/bridge2ai>. The Common Fund is managed by the Office of Strategic Coordination within DPCPSI.

²⁵Bélisle-Pipon J-C, Powell M, English R, Malo M-F, Ravitsky V, Bridge2AI-Voice Consortium, Bensoussan Y, *Stakeholder perspectives on ethical and trustworthy voice AI in health care*, Digital Health 10 (2024), <https://doi.org/10.1177/20552076241260407>.

²⁶Nebeker C, Bélisle-Pipon J-C, et al., *Ethical sourcing in the context of health data supply chain management: a value sensitive design approach*, JAMIA Open 8(5):ooaf101 (2025), <https://doi.org/10.1093/jamiaopen/ooaf101>.

²⁷Malo M-F, Bouhouita-Guermech S, Gallois H, Ravitsky V, Bridge2AI-Voice Consortium, Bélisle-Pipon J-C, *Using voice and speech data in healthcare: a scoping review of the ethical, legal and social implications*, Frontiers in Digital Health 8:1750111 (2026), <https://doi.org/10.3389/fdgth.2026.1750111>.