Comments of the World Privacy Forum
to the
European Data Protection Board
regarding
Guidelines on Processing of Personal Data for Scientific Research Purposes, Adopted 15 April, 2026 (version for public consultation)

25 June 2026

The World Privacy Forum (WPF) [1] appreciates the opportunity to comment on the EDPB's **Guidelines on Processing of Personal Data for Scientific Research Purposes, Adopted 15 April, 2026 (version for public consultation).**[2]

Founded in 2002, the World Privacy Forum is an independent public-interest research organization whose work concerns the privacy and governance of complex data ecosystems at scale, with particular depth in the governance of artificial intelligence and machine learning and in health data. Its research spans AI and machine-learning governance, biomedical and genomic privacy, and identity systems, including published research on the U.S. National Institutes of

[1] World Privacy Forum, https://worldprivacyforum.org/.

[2]European Data Protection Board, *Guidelines 1/2026 on the Processing of Personal Data for Scientific Research Purposes* (Version 1.0, adopted Apr. 15, 2026, public consultation until June 25, 2026), https://www.edpb.europa.eu/our-work-tools/documents/public-consultations/2026/guidelines12026-processing-personal-data-scientific_en.

Health's All of Us genomic biobank,[3] and draws on original work and policy engagement across 87 jurisdictions. The Forum's peer-reviewed field research on India's Aadhaar identity system, published in Springer Nature's *Health and Technology*,[4] was cited in the Supreme Court of India's landmark privacy decision.

The World Privacy Forum is pleased to comment on the EDPB's draft guidelines on processing of personal data for scientific research. As the EDPB is well aware, this subject is complex, presenting many layers of controversial and conflicting policy choices. Given enough time, our comments could be quite extensive. Luckily or not, time is short and we can only manage to provide a few brief comments in the time available to us.

I. Artificial Intelligence

The consequences of artificial intelligence (AI) for many human endeavors are being actively explored all around the world. Work on the interface between AI and data protection is ongoing widely. The draft guidelines do not address the consequences for data protection in a world where AI is a tool and a resource commonly available to scientific researchers and to others in related and unrelated activities. This is a vast subject worthy of more attention by data protection authorities, and the issues extend beyond the boundaries of the guidelines.

Nevertheless, we do not believe that it is advisable at this time for standards governing scientific research and personal data to ignore AI. At a minimum, we suggest that the guidelines should address the possibility (and the likelihood) that personal data provided to scientific researchers might be analyzed by or otherwise added to all-purpose large language models and will thus escape the intent and the controls of the guidelines. Scientific researchers should be expressly advised about this possibility and should be directed to prevent the "escape" of personal data to AI resources that are not expressly limited to appropriate scientific research. Once an individual's health data enters a large language model, that data may not only affect the individual, but also the individual's family, and in some cases that impact may last in perpetuity. It is also likely that data once disclosed in this manner will not be retrievable or able to be removed. These possibilities, together with the inference risks discussed below, may be the AI issues that warrant the highest degree of concern at this time.

We note that there is a substantial technical literature regarding Membership Inference Attacks, or MIA, that speaks to human subject research, and in particular, to genomic research. The essence of this body of work is that it proves that the participation of an identifiable individual in a research dataset can be inferred without access to any individual-level record — from

---

[3]Robert Gellman & Pam Dixon, *Privacy, the Precision Medicine Initiative, & the All of Us Research Program: Will Any Legal Protections Apply?* World Privacy Forum (2017), https://worldprivacyforum.org/posts/report-privacy-the-precision-medicine-initiative-and-all-of-us-research-program-will-any-legal-protections-apply/.

[4]Pam Dixon, *A Failure to Do No Harm: India's Aadhaar Biometric ID Program and Its Inability to Protect Privacy in Relation to Measures in Europe and the U.S.*, Health and Technology (Springer Nature, 2017), DOI 10.1007/s12553-017-0202-6, open access at https://techscience.org/a/2017082901/.

aggregate allele frequencies or summary statistics, from the yes-or-no responses of an allele-query service, or from a trained statistical or machine-learning model alone — and that this inference remains possible after the data have been de-identified and aggregated by the methods ordinarily relied upon to protect research participants. Research cohorts are commonly assembled by phenotype: the data that establishes that a person was a member of a cohort can be the same data that discloses the condition, disease, or heritage that defined it for that person and, because genomes are shared, for that person's relatives as well.

The same class of techniques has been used to recover individual training records and personal identifiers from a deployed large language model simply by querying it, which makes plain that the "escape" of personal data is not a hypothetical risk. The work of Homer et al. (2008)[5] demonstrated inference attacks in the genomic context that were consequential enough that the National Institutes of Health and the Wellcome Trust promptly withdrew aggregate genomic data from open access.[6] Shringarpure and Bustamante (2015)[7] published data on genomic data-sharing beacons, Shokri et al. (2017)[8] published data on trained machine-learning models generally, and Carlini et al. (2021)[9] published research on large language models in particular. This is a large literature demonstrating a meaningful problem, and significant for your undertaking.

We acknowledge that placing limits on personal data collected and compiled for scientific research may have unintended and unwelcome consequences for other scientific research activities. Tradeoffs are inherent in any data protection restrictions. We are all used to recognizing these types of problems and finding appropriate balances and compromises. We raise the concern about the possibility of the "escape" of personal data collected for research into general purpose AI systems because it is so huge and so consequential.

II. Groups

The focus of the guidelines on the role of individual consent does not take into account the complexity of scientific research on data protection matters that relate to groups of individuals. In some cases, there will be a group interest in collective privacy that should not be ignored. By collective privacy we mean "those privacy interests held by or applicable to a definable group of people — interests that include properties not reducible to any individual member apart from the

---

[5]Nils Homer et al., *Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays*, 4 PLoS Genetics e1000167 (2008).

[6]Elias A. Zerhouni & Elizabeth G. Nabel, *Protecting Aggregate Genomic Data*, 322 Science 44 (2008).

[7]Suyash S. Shringarpure & Carlos D. Bustamante, *Privacy Risks from Genomic Data-Sharing Beacons*, 97 American Journal of Human Genetics 631 (2015).

[8]Reza Shokri, Marco Stronati, Congzheng Song & Vitaly Shmatikov, *Membership Inference Attacks Against Machine Learning Models*, 2017 IEEE Symposium on Security & Privacy 3 (2017).

[9]Nicholas Carlini et al., *Extracting Training Data from Large Language Models*, 30th USENIX Security Symposium 2633 (2021).

group."[10] The previous section regarding challenges with AI and ML systems illustrates collective consequences: the privacy harm in scientific research, particularly in the realm of genomic databases, begins with collective harm – it attaches to group membership before it attaches to individual disclosure. The individual right of privacy cannot reach the group as such. The problem we describe here briefly is not made any easier by the difficulty of defining what constitutes a group that gives rise to a collective interest in data protection. We can all understand that tribal communities in many areas around the world are likely to have collective interests in data protection and other matters. These interests may be deeply rooted in socio-legal norms. In at least one case, the Māori, these interests are also rooted in a Māori Data Governance Model[11] and a separate formal agreement with the government of New Zealand via the Algorithm Charter for Aotearoa New Zealand.[12]

Others with a common genetic heritage or characteristic may be a group with collective interests of a different sort that warrants recognition, but that group may lack any formal recognition, governance structure, or communications mechanism. The link need not be heritable: families with a history of Huntington's Disease are linked by an inherited marker, while people identifiable through the somatic markers of a rare or orphan disease such as melorheostosis are linked by a marker no one passed down and no one can pass on, but the link is nevertheless definitive. Other groups that can be theoretically identified may have no meaningful common interests that give rise to data protection or other concerns. Consider, as an example, everyone who borrowed a book from a public library anywhere in the world on June 1, 2026. This hypothetical group may have no meaningful common interest that warrants attention or protection.

This aspect of data protection is largely unaddressed in formal data protection policy documents, but it will grow increasingly pressing in the future, especially as AI, now and increasingly, allows for the connection of information and individuals in ways that were at one time impossible or impractical. There is another issue underlying the importance of both addressing groups in the governance context: it is an essential technical fact that machine learning systems generate and act on groups as a core function of how they operate. Grouping, or clustering, is the air that AI/ML systems breathe, and this clustering is subject to Kleinberg's Impossibility Theorem[13] and the laws of mathematics. These AI/ML groups are generated without limit, and faster than our current frameworks can follow.

---

[10]Pam Dixon, *Group Privacy, Data, and AI: Collective Forms of Privacy and Its Relationship to Technology and Policy Frameworks*, J. Personal Data Protection L., No. 1–2, at 68 (Dec. 2025).

[11]Tahu Kukutai et al., *Māori Data Governance Model*, Te Kāhui Raraunga (2023), https://www.kahuiraraunga.io/_files/ugd/b8e45c_803c03ffe532414183afcd8b9ced10dc.pdf.

[12]*Algorithm Charter for Aotearoa New Zealand*, Stats NZ (July 2020), https://www.data.govt.nz/assets/data-ethics/algorithm/Algorithm-Charter-2020_Final-English-1.pdf.

[13]Jon Kleinberg, *An Impossibility Theorem for Clustering*, in Advances in Neural Information Processing Systems 15, at 463 (2002), https://papers.nips.cc/paper/2002/hash/43e4e6a6f341e00671e123714de019a8-Abstract.html.

While there is no simple approach to the "group" privacy problem, recognition is a first and important step. It would be appropriate for the guidelines to advise scientific researchers to be aware how their work may affect groups of individuals with collective interests that warrant attention. Groups recognized under existing laws as deserving of recognition and protections provide a starting point for identification. In the research context, that starting point already exists. The Council of Europe's Recommendation CM/Rec(2016)6 on research on biological materials of human origin provides that "possible risks for any individual in the same group as the person from whom biological materials have been removed should also be taken into consideration," and requires that "appropriate measures should be taken, in the full range of research activities, to prevent discrimination against, and to minimise the likelihood of stigmatisation of, any person, family or group."[14] The OECD's Recommendation on Human Biobanks and Genetic Research Databases makes the same recognition, noting that an association between a heritage and a particular disease may bring repercussions for families and communities who never participated.[15] An affirmative national model exists as well: the aforementioned Māori Data Governance Model articulates collective privacy expressly, including for algorithmic and AI processing, with the Algorithm Charter for Aotearoa New Zealand as a separate, related government instrument.

III. Oversight

We do not have any immediate solutions to the problems that we identify here. We only offer the notion that more attention and oversight may be one response. Existing institutions, like data protection authorities, may play an appropriate role in raising awareness of these issues and in finding responses. There may be a need for a more formal organization focused on scientific research that can conduct continuing oversight, address regulatory conflicts and differences, help scientific researchers identify and resolve specific issues that arise in research activities, and accept appeals from those individuals and groups that may be affected by the collection, use, and sharing of personal data for scientific research.

IV. International Differences in Approach

We observe that there are fundamental differences between the US and EU approaches that will make international cooperation in scientific research more difficult. It is probable that there are differences from subtle to significant across a variety of regions and jurisdictions. We suggest that the Board (or perhaps some organization designated by the Board) needs to take a closer look at these regional and national differences and suggest ways that scientific researchers can minimize administrative and governance conflicts while maximizing legitimate scientific

---

[14]Recommendation CM/Rec(2016)6 of the Committee of Ministers to Member States on Research on Biological Materials of Human Origin, Appendix (Guidelines), arts. 4(3), 5 (adopted by the Committee of Ministers on May 11, 2016).

[15]Recommendation of the Council on Human Biobanks and Genetic Research Databases, OECD/LEGAL/0375, Annotations ¶ 9 (adopted Oct. 22, 2009).

cooperation. These problems are not, of course, entirely new. However, given the presence of AI in the research ecosystems now, there is more urgency to find multi-jurisdictional agreement on approaches.

Thank you for the opportunity to provide these comments. We would welcome the opportunity to discuss these issues with you further.

Respectfully submitted,

Pam Dixon
Founder and Executive Director
World Privacy Forum