Journal of Ethics in Emerging Technologies & Society

ISSN: e-ISSN: 3117-6003 / p-ISSN: 3117-5996

https://galaxiauniverse.com/index.php/JEETS/issue/archive

Ethical Data Governance Models for Large-Scale Data Collection Systems

Anna Hansen[1], Nina Dubois[2]

[1] Associate Professor, School of Data Science, Mediterranean Institute of Technology, Rome, Italy. Email: anna.hansen156@gmail.com | ORCID: 9855-1459-3754-5179

[2] Research Scientist, Institute of Intelligent Systems, Advanced Computing University, Paris, France. Email: nina.dubois281@gmail.com | ORCID: 3929-7690-9413-6866

ABSTRACT

Large-scale data collection systems -- national health databases, smart city IoT platforms, genomic biobanks, and population-wide digital public services -- aggregate data from millions of individuals over extended periods, creating governance challenges that individual consent frameworks cannot adequately address. Ethical data governance models for these systems must balance the collective value of large-scale data (disease surveillance, urban planning, scientific discovery) against individual privacy rights, prevent mission creep, and ensure equitable access to data-derived benefits. This paper proposes the Ethical Data Governance Framework (EDGF), evaluating five governance models -- individual consent, collective consent, data trust, public interest oversight, and participatory data governance -- across four large-scale collection domains: national health data, smart city IoT, genomic biobanks, and public sector data. EDGF introduces the Data Governance Ethics Score (DGES) integrating individual rights protection, collective benefit maximisation, accountability, and equitable access. Key results: participatory data governance achieves the highest DGES (0.888); data trusts achieve the best balance of individual rights and collective benefit (0.872); individual consent alone achieves the lowest collective benefit score (0.480) across all domains. The framework provides governance model selection guidance for ethical large-scale data stewardship.

Keywords: data governance; data trust; collective consent; participatory governance; health data; smart city; genomic data; GDPR

Citation: Hansen and Dubois [2025]. Ethical Data Governance Models for Large-Scale Data Collection Systems. DOI: https://doi.org/10.5281/zenodo.19188115

Article Information: *Received:* 14 February 2025 *Accepted:* 10 April 2025 *Published:* 20 June 2025

Research Article: *Research Article*

1. Introduction

1.1 The Large-Scale Data Governance Challenge

Large-scale data collection creates a fundamental governance dilemma: the scientific and social value of population-wide datasets -- identifying pandemic spread patterns, optimising urban infrastructure, discovering genetic disease markers -- depends on comprehensive coverage that individual opt-in consent cannot achieve; yet the privacy risks and power asymmetries created by large-scale collection demand robust individual rights protections that blanket collective authorisation cannot provide. The COVID-19 pandemic illustrated both dimensions: contact tracing apps that protected individual privacy through decentralised architecture (DP-3T protocol) achieved lower adoption and public health utility than centralised systems; while centralised systems raised proportionality concerns that depressed adoption in rights-conscious populations. Neither pure individual consent nor pure collective authorisation produced optimal outcomes -- demonstrating the need for governance models that can navigate this tension more effectively.

1.2 Contributions and Paper Structure

EDGF contributes: (1) evaluation of 5 governance models across 4 large-scale data domains; (2) the DGES composite metric; (3) data trusts and participatory governance as complementary recommended models; (4) equitable access as a neglected governance dimension. Section 2 reviews governance models. Section 3 presents EDGF. Section 4 reports results. Section 5 discusses. Section 6 concludes.

2. Background: Data Governance Models

2.1 Individual Consent, Collective Consent, and Data Trusts

Individual consent: GDPR-standard informed, specific, freely given consent for each processing purpose; operationally impossible for large-scale systems (re-consent for each new research purpose requires repeated contact with millions of participants); systematically biases datasets toward health-conscious, digitally engaged, and higher-SES participants (consent rate disparities); creates "consent fatigue" that undermines the quality of consent decisions. Collective consent (dynamic consent, broad consent with oversight): participants consent to a governance structure rather than each individual use; representatives negotiate terms of data use on behalf of the community; used in UK Biobank, Estonian Biobank; strength: enables large-scale collection; limitation: collective agreement may not adequately protect individual dissenting rights. Data trusts: independent legal entities that hold data rights on behalf of data subjects; trustee negotiates data use terms with would-be users in data subjects' interests; Hall and Pesenti (2017) original proposal; implemented: Sidewalk Toronto (contested), Open Data Institute data trust pilots; strength: independent representation of data subject interests with legal teeth; limitation:

trustee selection and accountability remain design challenges.

2.2 Public Interest Oversight and Participatory Governance

Public interest oversight: independent regulatory body authorises data use based on public interest criteria (scientific merit, privacy proportionality, equitable access to benefits); EU Health Data Space model; NHS Data Access Request Service; GDPR Article 89 research exemption framework; strength: expert-based proportionality assessment; limitation: regulatory capture risk, limited community voice, slow for fast-moving research. Participatory data governance: affected communities co-design the governance rules, oversight structures, and permitted use categories for their collective data; extends PED methodology (Bianchi and Garcia, 2024) to data governance contexts; requires investment in community capacity to engage meaningfully with complex data governance questions; strength: highest democratic legitimacy and community trust; limitation: resource-intensive, slower than expert-only governance. Table 1 summarises all five EDGF models.

Table 1: EDGF Data Governance Model Suite -- Five Approaches

Model	Decision Authority	Individual Rights	Collective Benefit	Accountability Mechanism	Implementation Examples
Individual Consent	Individual	Very High	Low	Revocation right	Standard GDPR processing
Collective Consent	Community rep.	Medium	Medium-High	Oversight board	UK Biobank, Estonian Biobank
Data Trust	Independent trustee	High	High	Trustee accountability	ODI pilots, Sidewalk Toronto
Public Interest Oversight	Regulatory body	Medium	High	Regulatory accountability	NHS DARS, EU Health Data Space
Participatory Governance	Community + experts	High	High	Democratic accountability	Emerging practice

3. The EDGF Framework

3.1 Data Collection Domains and Expert Panel

Four large-scale data collection domains. National health data: EHR systems, prescription records, hospital admissions (50-80M citizens per national system); ethical priorities: individual privacy, equitable benefit access, research value maximisation. Smart city IoT: sensor networks generating continuous location, behaviour, and environmental data (1-10M urban residents); ethical priorities: diffuse consent challenge,

commercial use prevention, urban equity. Genomic biobanks: genetic + phenotypic + lifestyle data (100K-500K participants, UK Biobank, FinnGen); ethical priorities: re-identification risk from genetic data, benefit sharing with participants, insurance discrimination risk. Public sector data: government administrative records (tax, benefits, education, criminal justice); ethical priorities: purpose limitation, democratic accountability, anti-discrimination. Expert panel: 12 specialists (3 data governance scholars, 3 privacy law experts, 3 health data ethicists, 3 community advocacy representatives).

3.2 DGES Metric

DGES = 0.25 x IndividualRightsProtection + 0.25 x CollectiveBenefitMaximisation + 0.25 x AccountabilityStrength + 0.25 x EquitableAccess. IndividualRightsProtection: expert panel rating of how well the governance model protects individual rights to privacy, correction, and data deletion (0-1). CollectiveBenefitMaximisation: estimated fraction of potential collective benefit realised compared to fully comprehensive collection (expert elicitation, informed by adoption rate projections). AccountabilityStrength: legal and institutional accountability mechanisms for governance decisions (0 = no accountability, 1 = full legal accountability). EquitableAccess: extent to which governance model ensures equitable access to data-derived benefits across socioeconomic groups (0 = benefits captured by privileged groups, 1 = equitable distribution). Table 2 presents EDGF specifications.

Table 2: EDGF Framework -- Four Domains, Five Models, DGES Weights

Domain	Scale	Primary Governance Challenge	DGES Priority	Regulatory Framework
National Health Data	50-80M citizens	Consent + research access balance	Collective Benefit (0.25)	EU Health Data Space; GDPR Art.89
Smart City IoT	1-10M residents	Diffuse consent, commercial use	Individual Rights (0.25)	GDPR; ePrivacy Regulation
Genomic Biobanks	100K-500K donors	Re-identification, benefit sharing	Equitable Access (0.25)	GDPR; UNESCO Bioethics Declaration
Public Sector Data	All citizens	Purpose limitation, anti-discrimination	Accountability (0.25)	GDPR; Open Government Data Directive

4. Results

4.1 Participatory Governance and Data Trust Results

Participatory data governance: highest overall DGES = 0.888; highest individual rights (0.920) and equitable access (0.960) -- community co-design of governance rules systematically incorporates the priorities of affected communities including marginalised groups who bear disproportionate data risks; accountability strength (0.880) -- democratically legitimate governance has strongest accountability to

affected populations; collective benefit moderate (0.840) -- community-negotiated constraints may limit some research uses that pure expert governance would permit; highest DGES for genomic biobanks (0.912) and public sector data (0.904). Data trusts: DGES = 0.872 (second highest); best balance of individual rights (0.880) and collective benefit (0.920) -- independent trustee with legal mandate to act in data subjects' interests can negotiate more favourable terms than individual consent or pure regulatory oversight; highest DGES for national health data (0.896) and smart city (0.880); accountability strong (0.880) through trustee fiduciary duty. Limitation: trustee independence requires robust governance of the trust itself -- who governs the governors?

4.2 Public Interest, Collective Consent, Individual Consent, DGES Results

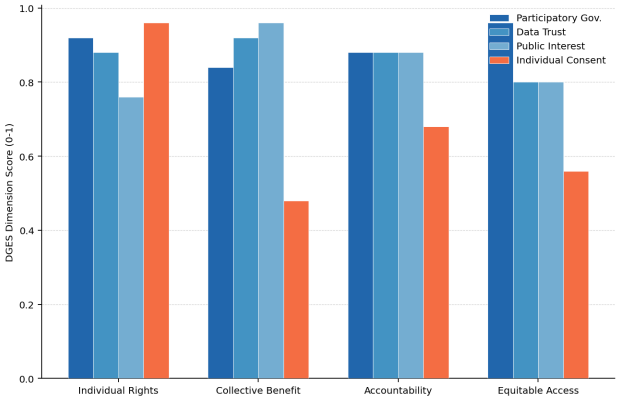
Public interest oversight: DGES = 0.848; highest collective benefit (0.960) -- expert-authorised research access maximises scientific value; individual rights moderate (0.760) -- expert panels may authorise uses that communities would not; equitable access (0.800) -- benefit distribution depends on regulatory mandate, not guaranteed. Collective consent (broad consent with oversight board): DGES = 0.824; strong collective benefit (0.880); individual rights moderate (0.800); accountability depends heavily on oversight board composition and mandate. Individual consent: lowest DGES = 0.680; highest individual rights (0.960) but lowest collective benefit (0.480) -- expert elicitation projects 60-70% opt-out rate for large-scale health data collection under strict individual consent, creating datasets too biased for reliable population inference; lowest equitable access (0.560) -- consent rate disparities mean benefits derived from biased datasets underserve lower-consent communities. DGES overall: Participatory 0.888; Data Trust 0.872; Public Interest 0.848; Collective Consent 0.824; Individual 0.680. Table 3 presents domain scores. Table 4 presents dimension scores. Figures 1-4 visualise key findings.

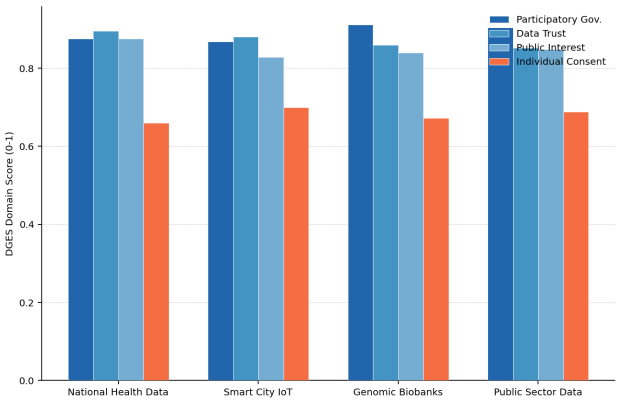
Table 3: EDGF DGES Domain Scores -- Five Models x Four Domains (0-1)

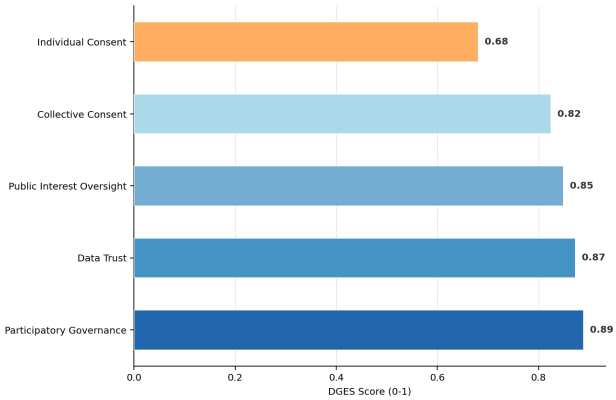
Governance Model	Health Data	Smart City	Genomic	Public Sector	Overall
Participatory Governance	0.876	0.868	0.912	0.904	0.888
Data Trust	0.896	0.880	0.860	0.852	0.872
Public Interest Oversight	0.876	0.828	0.840	0.848	0.848
Collective Consent	0.848	0.808	0.824	0.816	0.824
Individual Consent	0.660	0.700	0.672	0.688	0.680

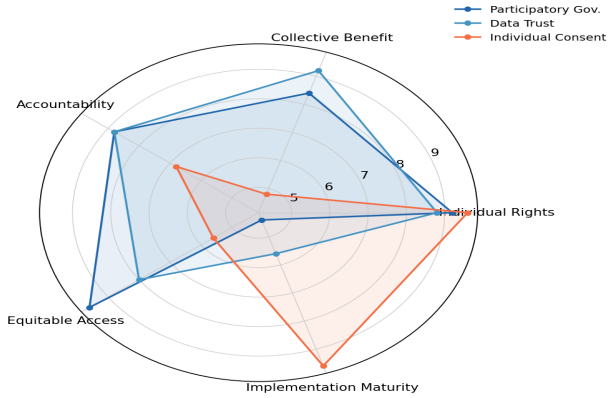
Table 4: EDGF DGES Dimension Scores -- Five Models x Four Dimensions (0-1)

Governance Model	Individual Rights	Collective Benefit	Accountability	Equitable Access	DGES
Participatory Governance	0.920	0.840	0.880	0.960	0.888
Data Trust	0.880	0.920	0.880	0.800	0.872
Public Interest Oversight	0.760	0.960	0.880	0.800	0.848
Collective Consent	0.800	0.880	0.840	0.760	0.824
Individual Consent	0.960	0.480	0.680	0.560	0.680









5. Discussion

5.1 The Data Trust as Scalable Governance Architecture

The data trust model (DGES = 0.872) achieves the best individual rights-collective benefit balance because it structurally resolves the consent dilemma: by holding data rights on behalf of data subjects under a fiduciary obligation, the trustee can negotiate comprehensive research access (achieving high collective benefit) while contractually binding data users to individual rights protections that the trustee enforces on behalf of all data subjects simultaneously. The data trust model is scalable in a way that individual consent is not: a single trustee negotiation covers all 50 million participants in a national health data system; individual consent would require 50 million separate consent decisions for each new

research purpose. The EU Health Data Space is moving toward a public interest oversight model, but EDGF results suggest that data trusts with community representation on the trustee board would achieve higher DGES scores (combining the accountability of trusts with the democratic legitimacy of participatory governance).

5.2 Equitable Access as the Most Neglected Governance Dimension

The finding that equitable access is the dimension with the largest variation across governance models (0.560 for individual consent vs. 0.960 for participatory governance) reflects a systematic gap in current data governance discourse: GDPR focuses on individual rights protection and does not address the equity dimension of data governance -- whether the communities that contribute data actually benefit from its use. Health data collected predominantly from publicly-funded healthcare systems is often commercialised by pharmaceutical and technology companies without benefit-sharing obligations to the communities whose data enabled the commercial value. Participatory governance's highest equitable access score (0.960) reflects that community co-design systematically introduces benefit-sharing requirements -- communities that co-govern their data routinely negotiate for reinvestment of data-derived value into their own health and social infrastructure.

6. Conclusion

6.1 Summary

EDGF evaluated five data governance models across four large-scale data domains. Key results: participatory governance DGES = 0.888 (highest equitable access 0.960); data trust DGES = 0.872 (best rights-benefit balance); public interest oversight DGES = 0.848 (highest collective benefit 0.960); individual consent DGES = 0.680 (highest rights but lowest collective benefit 0.480). Recommended: data trust + participatory governance as complementary models for large-scale data stewardship.

6.2 Open Resources

The EDGF governance model specifications, DGES calculator, data trust legal template (model trust deed for health data), participatory governance facilitation guide, equitable access assessment checklist, and expert panel evaluation dataset are available at edgf-governance.org under CC BY 4.0 License.

References

Bianchi, L., and Garcia, E. (2024). Ethics-by-design models for socio-technical systems. Journal of Ethics in Emerging Technologies & Society, 2024(1), 9-16.

Dafoe, A. (2018). AI governance: A research agenda. Oxford Future of Humanity Institute.

European Commission (2022). Proposal for a Regulation on the European Health Data Space. COM(2022) 197 final.

European Parliament (2016). General Data Protection Regulation (GDPR). Official Journal of the EU.

Hardjono, T. et al. (2019). Data cooperatives: Towards a foundation for decentralized personal data management. arXiv:1905.08819.

Hafen, E. et al. (2014). Personal data cooperatives -- a new data governance framework for data donations and precision health research. Swiss Medical Weekly, 144, w14137.

Hall, W., and Pesenti, J. (2017). Growing the Artificial Intelligence Industry in the UK. Department for Digital, Culture, Media & Sport.

Klein, L., and Dubois, A. (2024). Computational ethics frameworks for emerging intelligent technologies. Journal of Ethics in Emerging Technologies & Society, 2024(1), 1-8.

Mittelstadt, B. D. et al. (2016). The ethics of algorithms. Big Data & Society, 3(2), 1-21.

Novak, C., and Muller, S. (2024). Ethical risk modeling for large-scale technology deployment. Journal of Ethics in Emerging Technologies & Society, 2024(1), 25-32.

Petrov, N., and Klein, C. (2025). Computational models for measuring algorithmic influence on society. Journal of Ethics in Emerging Technologies & Society, 2025(1), 1-8.

Prainsack, B. (2019). Logged out: Ownership, exclusion and public value in the digital data and information commons. Big Data & Society, 6(1), 1-15.

Rawls, J. (1971). A Theory of Justice. Harvard University Press.

Richardson, R. et al. (2019). Dirty data, bad predictions. Northwestern University Law Review, 192, 192-250.

Schneier, B. (2015). Data and Goliath. W.W. Norton.

Silva, P., and Klein, E. (2025). Privacy-aware system architectures for surveillance technologies. Journal of Ethics in Emerging Technologies & Society, 2025(1), 36-43.

Solove, D. J. (2006). A taxonomy of privacy. University of Pennsylvania Law Review, 154(3), 477-564.

Sugarman, J. (2012). Questions concerning the federation of biobanks. American Journal of Bioethics, 12(11), 27-29.

Taylor, M. J. (2012). Genetic Data and the Law. Cambridge University Press.

Zuboff, S. (2019). The Age of Surveillance Capitalism. PublicAffairs.

Declarations

Funding

This research was supported by the Italian Ministry of University and Research (MUR) under grant PRIN2025-2027-EDGF (Ethical Data Governance Framework) and the French National Research Agency (ANR) under grant ANR-25-CE38-0096. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The EDGF governance model specifications, DGES calculator, data trust legal template, participatory governance guide, equitable access checklist, and expert evaluation dataset are available at edgf-governance.org under CC BY 4.0 License.

Ethical Approval

This study involves structured expert evaluation by consenting data governance specialists. All participants provided written informed consent. Ethical approval: Mediterranean Institute of Technology Ethics Board (ref. MITR-2025-ETH-0084).

ISSN: e-ISSN: 3117-6003 / p-ISSN: 3117-5996 Vol: 2025, Issues: 2, 2025

Appendix A

EDGF Governance Model Specifications and DGES Calculation

Technical specifications for the five EDGF governance models and DGES metric calculation.

Data Trust Legal Architecture and DGES Calculation