JUST TECH

The Data We Don't See: Training Sets, Cultural Heritage, and the New Geopolitics of AI

By: Mika (Jaeyun) Noh

LAW & ETHICS

Noh, Mika (Jaeyun). "The Data We Don't See: Training Sets, Cultural Heritage, and the New Geopolitics of AI." Just Tech. Social Science Research Council. October 7, 2026. DOI: https://doi.org/10.35650/JT.3105.d.2026.

This essay is adapted from sections of Curating Intelligence: The Playbook for Art, Governance, and the New Global Economy*, published by ARTLAKE Inc.*

While public discourse focuses on semiconductors, cloud infrastructure, and large language models, it often ignores the importance of memory. Yet in the age of artificial intelligence, these are increasingly interconnected conversations. The datasets on which AI systems are trained are not neutral archives. They are selective reconstructions of the world, encoding whose knowledge counts, whose aesthetics are legible, whose histories are preserved with fidelity, and whose are flattened into noise. To control the training data is to shape the cognitive baseline of machine intelligence.

Governments, platforms, and communities are already competing on the terrain of cultural heritage and creation ecosystem. However, that competition remains undertheorized in mainstream AI policy discourse. The societies that are digitizing their archives, governing their training data, and building rights frameworks for creative workers are making decisions whose consequences will outlast any individual model or regulatory cycle. The societies that are not are making a different decision: to allow that architecture to be built for them, on terms set by others, encoding assumptions they did not choose.

In 2024 and 2025, the MIT Data Provenance Initiative completed a sweeping audit of more than 1,800 datasets used in training large-scale AI systems. Their findings were stark: Licenses were miscategorized in over half of all cases, and critical attribution information was entirely absent from more than 70

percent of analyzed datasets.[1] Researchers termed this phenomenon “license laundering”—the systematic obscuring of provenance chains that allows copyrighted and culturally sensitive material to be incorporated into training corpora without legal or ethical accountability.[2]

License laundering is not primarily a technical failure. It is a governance failure—the predictable consequence of a period in which the computational benefits of large datasets were pursued without corresponding investment in the rights frameworks that would govern their use. The artists, archivists, and cultural communities whose work entered these datasets were not consulted. Their contributions were not attributed. The value generated from their creativity accrued elsewhere.

The consequences are neither abstract nor evenly distributed. What is at stake is not merely accurate representation in a demographic sense or financial compensation, though these are also important. It is the question of which cultural frameworks machine intelligence learns to treat as universal.

Culture as Infrastructure, Not Ornament

South Korea’s K-wave—the global diffusion of South Korean pop music, drama, cinema, fashion, and food—is often told as a story of soft power. But beneath that narrative is a more important development: the emergence of culture as soft infrastructure.

The distinction matters. Soft power, as Joseph Nye theorized it, describes the capacity to attract and persuade through cultural appeal. Soft infrastructure is something more complex: the technical systems, data architectures, and governance frameworks that determine whether cultural products can circulate, be discovered, be attributed, and accumulate value in a world organized around algorithmic mediation. Soft power is an effect. Soft infrastructure is the condition of its production.

South Korea’s recent cultural prominence has been a data story. TikTok’s recommendation algorithm amplified K-pop’s global reach by an estimated 400 percent between 2020 and 2023—not through promotional expenditure but through cultural optimization, the capacity of algorithmic systems to identify and surface the formal features of content that drive engagement.[3] K-culture exports reached $13 billion in 2024, surpassing semiconductor exports.[4] While significant, this success also reveals a structural vulnerability: When cultural circulation depends on algorithmic infrastructure owned by global platforms, the terms of that circulation are set by platform owners, not cultural producers.

Algorithms increasingly determine which South Korean stories become globally legible—that is, which cultural products are formatted in ways that allow transnational platforms to classify, compare, rank, and optimize them for global audiences. Playlist engines, for example, can influence which K-pop sounds are treated as “export-ready” and therefore more readily surfaced to international listeners. AI-optimized trailers and auto-translated lyrics compress complex cultural contexts into metadata, embeddings, and engagement metrics that algorithms can act on. Cultural diplomacy has migrated from ministries and artists to data infrastructures headquartered in California. The nation exports cultural products; the architecture that determines its value and circulation is owned elsewhere.

The deeper risk is not replacement but hollowing. An AI system trained on K-pop can generate K-pop-adjacent content at scale. But it cannot generate the specific historical, linguistic, and emotional textures that make Korean cultural production distinctive. Those textures ought to be captured in high-fidelity, governed, sovereign datasets whose governance reflects the agency of the cultural producers themselves—artists, writers, archivists, and communities—through collective institutions such as data trusts, cooperatives, and commons-based licensing regimes. Otherwise, "Korean control" risks reproducing the same extraction, only under domestic elites. Without these protections, those textures remain invisible to the machines that will increasingly shape what the world encounters as "Korean." The danger, in short, is that AI will outperform Korean culture in circulation while eroding it in substance, turning cultural labor into training data and national identity into a style preset.

Heritage Data as Geopolitical Capital

The realization that cultural data is geopolitical capital is globally reshaping national strategies in ways that rarely receive adequate attention in AI policy discussions.

South Korea has determined that cultural heritage, once digitized and governed on South Korean terms, becomes exportable soft infrastructure. Heritage data is no longer static; it is an active resource that feeds production design, gaming, tourism, and AI training. Through South Korea's Digital Cultural Heritage Source Resource Distribution Project, extensive collections of 3D assets are now available for secondary use across industries. The loop is self-reinforcing: Governed cultural data generates economic value, which funds further digitization, which expands the governed archive.

Consider what South Korea's Electronics and Telecommunications Research Institute (ETRI) has built: a digital infrastructure combining large language models with geospatial data to translate and classify materials from classical Korean and Hanmun scripts—the historical literary language of the Korean court. Their aim is to create a "One Heritage Registry": a unified, computable database where a dancer's motion-captured movement, a temple's digitized roof beam, and an archived royal decree coexist within a single interoperable architecture.

By 2025, South Korea had established what amounts to a cultural data export strategy, providing high-precision 3D digital assets to European institutions and extending digital development assistance to safeguard heritage across the ASEAN region. What was once cultural diplomacy—the export of K-drama, K-food, K-fashion—has evolved into something structurally more durable: the export of the technical frameworks through which culture is stored, classified, and made accessible to intelligent systems.

The geopolitical stakes are significant. If South Korean-defined technical standards become the global benchmark for heritage digitization, then AI systems trained on those standards will carry Korean cultural assumptions as part of their baseline. The alternative is that AI systems learn from heritage data digitized and classified according to frameworks developed by institutions with the longest archival traditions and the deepest computational resources. Those institutions are overwhelmingly in the Global North. Whoever defines the standards by which cultural heritage is made machine-readable ultimately defines how the future remembers the past.

Korea’s Fourth Way

International AI governance discourse is typically organized around three dominant frameworks. The European Union’s AI Act represents a rights-based precautionary model: comprehensive regulation, categorical risk classification, and robust enforcement. The United States has pursued a market-driven adaptive approach: principles rather than binding rules, with sectoral interventions reserved for specific harms. China’s model emphasizes state-led developmental governance: AI as a strategic national capability, accelerated and aligned with party values at the training and deployment levels.

South Korea has pursued a Fourth Way.

The AI Basic Act, passed in 2025 and going into effect in January 2026, established a risk-based regulatory architecture that shares features with the EU approach—transparency requirements, impact assessments, record-keeping obligations for high-risk and generative systems—while explicitly framing these obligations as enabling rather than constraining cultural production. The act was designed as a foundational statute rather than a sector-specific rulebook, allowing downstream application in the arts and content industries where questions of authorship, training data provenance, and algorithmic influence are structurally unavoidable.

What distinguishes the South Korean approach is not any single policy instrument but the synchronization of regulatory, labor, and institutional mechanisms into what might be called a cultural operating system—a coherent governance architecture in which AI legislation, social insurance for artists, and museum-level institutional standards advance in parallel and reinforce one another.

South Korea’s Artist Employment Insurance, introduced in December 2020, formally recognizes artists as insured workers rather than treating them as informal or purely cultural participants. This gives artists a concrete economic stake in the governance of AI: If AI systems misattribute or undercompensate their work, the resulting loss of income can affect not only their earnings but also their access to employment-related protections and benefits. Artists, therefore, have a material interest in AI governance frameworks that protect attribution, remuneration, and other economic rights.

At the institutional level, the National Museum of Modern and Contemporary Art (MMCA) has positioned itself as a test bed for curatorial transparency by piloting AI exhibitions and smart-museum projects that explicitly disclose AI involvement, address rights-holder consent, and experiment with curator-controlled AI tools for interpretation. These practices make data sources, human–AI authorship, and the role of algorithmic processes more visible within the exhibition context. In this sense, MMCA provides an institutional example of how provenance, consent, and data-lineage questions can be incorporated into curatorial practice rather than treated solely as external regulatory concerns. As policy and museum discourse increasingly converge around provenance and data accountability, such practices point toward an emerging model in which curatorial transparency becomes part of professional practice rather than remaining only an ethical aspiration.

Where other systems hesitate—constrained by entrenched interests, jurisdictional fragmentation, or an

ideological commitment to either markets or precaution—South Korea legislates, tests, and iterates. Opacity, in this model, is not tradition. It is risk.

The Consent Architecture

The governance failure at the heart of the training data crisis is ultimately a consent failure. The legal frameworks that governed cultural production in the twentieth century—copyright, moral rights, attribution requirements—were not designed for a world in which the entire accessible archive of human creative output could be scraped, processed, and used to train systems whose outputs compete with the humans who produced them.

Addressing this failure requires building what I call a "consent architecture": a coordinated suite of legal, technical, and institutional mechanisms—such as licensing frameworks, provenance tracking, data trusts, and museum and industry standards—that together restore agency to cultural producers over how their work enters AI training contexts. The elements of this architecture are becoming clearer, even if their implementation remains uneven.

Governments have three core obligations. First, they should support the creation of licensed cultural archives: registries[5] for training data that operate on an explicit opt-in basis and provide a legal alternative to indiscriminate scraping. These archives need not be state controlled; they could be publicly mandated but independently governed through cultural data trusts, library and museum consortia, or collective management organizations acting on behalf of rights holders and relevant communities. Such arrangements could generate governed and attributed datasets while accommodating different forms of cultural ownership and stewardship. Second, governments should establish fee-based compensation models and downstream remuneration mechanisms for creators whose work is used in large-scale AI training. Third, they should develop legal frameworks that allow creators to set specific and enforceable limits on how their data is used, making attribution and consent meaningful rather than merely formal.

For cultural institutions, the corresponding obligations center on algorithmic stewardship: multidisciplinary ethics committees capable of reviewing AI projects for bias and rights compliance before deployment; transparency labeling that discloses when and how AI was used in the production or curation of works; and systematic audit of data lineage to ensure that the provenance of training materials respects the cultural and creative rights of their originators.[6]

South Korea's AI Basic Act provides a legislative template, although its implementation remains a work in progress. Separately, the Korea Copyright Act provides the substantive legal framework governing authorship and copyright protection. Article 2.1 of the Copyright Act specifies that a "work" must express human thoughts and emotions, thereby excluding purely AI-generated outputs from traditional copyright protection and establishing the principle that authorship requires human agency. The act's provisions for high-impact AI companies, such as major AI developers and providers whose systems are deployed in areas such as recruitment, credit assessment, healthcare, education, or public services, to identify and mitigate risks across the AI lifecycle create a regulatory infrastructure that can be extended to cultural

rights contexts as the jurisprudence matures.

What Is at Stake

A society that allows its cultural heritage to be incorporated into AI training corpora without attribution, compensation, or governance is not merely failing its artists. It is surrendering its cultural cognitive baseline—the specific textures of memory, aesthetics, and meaning that distinguish its machine intelligence from that produced by any other society. In an age when AI systems are becoming the primary mediators of cultural experience globally, that surrender has consequences that will compound across generations.

The first half of the twentieth century belonged to nations that mastered steel. The second half to those that mastered semiconductors. The twenty-first century will belong, in significant part, to those who can produce memory at scale—who can transform their cultural heritage into living, interoperable knowledge systems capable of being understood, used, and built upon by intelligent machines. This is not a metaphor for cultural vitality. It is a description of technical infrastructure.

The question that remains is political: whether the communities, institutions, and governments with the most to gain from culturally sovereign AI will act while the architecture is still being designed. Every dataset, every governance framework, and every curatorial decision made now is a choice about whose version of the past will train the intelligence of the future. The algorithms are being written. The question is whether culture will write them—or be written by them.

Footnotes

1 Shayne Longpre et al., "Consent in Crisis: The Rapid Decline of the AI Data Commons" (MIT Data Provenance Initiative, 2024), https://www.dataprovenance.org/Consent_in_Crisis.pdf.

2 James Jewitt et al., "Don't Trust the Label: License Laundering in AI Supply Chains," preprint, arXiv, July 22, 2026, https://doi.org/10.48550/arXiv.2607.20300 .

3 K-pop algorithm amplification figure (400 percent global increase, 2020–2023) attributed to platform analytics reporting on TikTok recommendation effects. Discussed in Mika Noh, "The Algorithm's Gaze vs. the Artist's Hesitation," chap. 1 in *Curating Intelligence* (ARTLAKE Inc., 2026).

4 K-culture export figure ($13 billion, 2024); see Noh, "The Algorithm's Gaze vs. the Artist's Hesitation."

5 The governance of historical and cultural materials requires a differentiated approach. For works still protected by copyright, permission should generally come from relevant rights holders. Public-domain status, however, does not necessarily resolve questions of ethical stewardship: Museums, archives, donor agreements, and source communities may establish contractual, institutional, or community-based protocols governing access, attribution, and the use of materials in AI training.

6 Recommendations on ethics committees, transparency labeling, and data lineage audit are drawn from Noh, "The Takedown Notice: Copyright, Liability, and the Data Minefield," chap. 5, and "Art Law as Global Benchmark: Korea's Transparency Act Model," chap. 8 in *Curating Intelligence*. Government-facing obligations—licensed cultural archives, compensation models, enforceable limits—are elaborated in Noh, "The Takedown Notice."