

Zbornik 29. mednarodne multikonference
INFORMACIJSKA DRUŽBA – IS 2026
Zvezek M

Proceedings of the 29th International Multiconference
INFORMATION SOCIETY – IS 2026
Volume M

Konferenca o umetniški inteligenci in pospeševanje
doseganja ciljev trajnostnega razvoja
Conference on Artificial Intelligence and Accelerating the Sustainable
Development Goals (AI4SDG)

Uredniki / Editors

Joao Pita Costa, Joao Paulo Vitor, Anja Polajner, Andrej Srakar

<https://is.ijs.si>

9. oktober 2026 / 9 October 2026
Ljubljana, Slovenia

Uredniki:

Joao Pita Costa
IRCAI, Ljubljana, Slovenia

Joao Paulo Veiga
CIAAM/USP, Sao Paulo, Brasil

Anja Polajner
IRCAI, Ljubljana, Slovenia

Andrej Srakar
IRCAI, Ljubljana, Slovenia

Založnik: Institut »Jožef Stefan«, Ljubljana
Priprava zbornika: Mitja Lasič, Vesna Lasič
Oblikovanje naslovnice: Vesna Lasič

Dostop do e-publikacije:
<https://library.ijs.si/Stacks/Proceedings/InformationSociety>

Ljubljana, oktober 2021

Informacijska družba
ISSN 2630-371X



PREDGOVOR MULTIKONFERENCI INFORMACIJSKA DRUŽBA 2026

29. mednarodna multikonferenca *Informacijska družba* se odvija v času izjemne rasti umetne inteligence, njenih aplikacij in vplivov na človeštvo. Vsako leto bolj vstopamo v novo dobo, v kateri generativna umetna inteligenca ter drugi inovativni pristopi oblikujejo poti k superinteligenci in singularnosti, ki bosta krojili prihodnost človeške civilizacije. Naša konferenca je tako hkrati tradicionalna znanstvena in akademsko odprta, pa tudi inkubator novih, pogumnih idej in pogledov.

Letošnja konferenca poleg umetne inteligence vključuje tudi razprave o perečih temah današnjega časa: ohranjanje okolja, demografski izzivi, zdravstvo in preobrazba družbenih struktur. Razvoj UI ponuja rešitve za številne sodobne izzive, kar poudarja pomen sodelovanja med raziskovalci, strokovnjaki in odločevalci pri oblikovanju trajnostnih strategij. Zavedamo se, da živimo v obdobju velikih sprememb, ko je ključno, da z inovativnimi pristopi in poglobljenim znanjem ustvarimo informacijsko družbo, ki bo varna, vključujoča in trajnostna.

V okviru multikonference smo letos združili dvanajst vsebinsko raznolikih srečanj, ki odražajo širino in globino informacijskih ved: od umetne inteligence v zdravstvu, demografskih in družinskih analiz, digitalne preobrazbe zdravstvene nege ter digitalne vključenosti v informacijski družbi, do raziskav na področju kognitivne znanosti, zdrave dolgoživosti ter vzgoje in izobraževanja v informacijski družbi. Pridružujejo se konference o legendah računalništva in informatike, prenosu tehnologij, mitih in resnicah o varovanju okolja, odkrivanju znanja in podatkovnih skladiščih ter seveda Slovenska konferenca o umetni inteligenci.

Poleg referatov bodo okrogle mize in delavnice omogočile poglobljeno izmenjavo mnenj, ki bo pomembno prispevala k oblikovanju prihodnje informacijske družbe. »Legende računalništva in informatike« predstavljajo domači »Hall of Fame« za izjemne posameznike s tega področja. Še naprej bomo spodbujali raziskovanje in razvoj, odličnost in sodelovanje; razširjeni referati bodo objavljeni v reviji *Informatica*, s podporo dolgoletne tradicije in v sodelovanju z akademskimi institucijami ter strokovnimi združenji, kot so ACM Slovenija, SLAIS, Slovensko društvo Informatika in Inženirska akademija Slovenije.

Vsako leto izberemo najbolj izstopajoče dosežke. Letos je nagrado *Michie-Turing* za izjemen življenjski prispevek k razvoju in promociji informacijske družbe prejel **Denis Trček**, priznanje za raziskovalni dosežek leta pa **Jure Leskovec**. »Informacijsko limono« za najmanj primerno informacijsko storitev je prejel informacijski sistem za dolgotrajno oskrbo. »Informacijsko jagodo« za najboljši sistem ali storitev v letih 2025/2026 pa je prejela elektronska osebna izkaznica. Iskrene čestitke vsem nagrajencem!

Naša vizija ostaja jasna: prepoznati, izkoristiti in oblikovati priložnosti, ki jih prinaša digitalna preobrazba, ter ustvariti informacijsko družbo, ki koristi vsem njenim članom. Vsem sodelujočim se zahvaljujemo za njihov prispevek — veseli nas, da bomo skupaj oblikovali dosežke, ki bodo krojili usodo Slovenije in sveta.

Mojca Ciglarič, predsednica programskega odbora
Matjaž Gams, predsednik organizacijskega odbora

FOREWORD TO THE MULTICONFERENCE INFORMATION SOCIETY 2026

The 29th International Multiconference *Information Society* takes place at a time of remarkable growth in artificial intelligence, its applications, and its impact on humanity. Each year we move further into a new era in which generative artificial intelligence and other innovative approaches are opening paths toward superintelligence and the singularity — two developments that will shape the future of human civilization. Our conference is at once a traditional scientific meeting, an academically open forum, and an incubator of new and bold ideas.

In addition to artificial intelligence, this year's conference also addresses pressing issues of our time: environmental protection, demographic challenges, healthcare, and the transformation of social structures. The development of AI offers solutions to many contemporary challenges and underscores the need for collaboration among researchers, professionals, and policymakers in shaping sustainable strategies. We are aware that we live in a period of profound change, when innovative approaches and deep knowledge are essential if we are to build an information society that is safe, inclusive, and sustainable.

This year's multiconference brings together twelve thematically diverse meetings that reflect the breadth and depth of the information sciences. They range from artificial intelligence in healthcare, demographic and family studies, the digital transformation of nursing, and digital inclusion in the information society, to research in cognitive science, healthy longevity, and education in the information society. They are joined by conferences on the Legends of Computing and Informatics, technology transfer, myths and truths of environmental protection, knowledge discovery and data warehouses, and, of course, the Slovenian Conference on Artificial Intelligence.

Alongside the papers, round tables and workshops will enable an in-depth exchange of views that will contribute substantially to shaping the future information society. *Legends of Computing and Informatics* serves as a national Hall of Fame for outstanding individuals in the field. We will continue to foster research and development, excellence, and collaboration. Extended papers will be published in the journal *Informatica*, building on a long-standing tradition and on cooperation with academic institutions and professional associations, including ACM Slovenia, SLAIS, the Slovenian Society Informatika, and the Slovenian Academy of Engineering.

Each year we single out the most distinctive achievements. This year the *Michie–Turing Award* for an outstanding lifetime contribution to the development and promotion of the information society was presented to **Denis Trček**, and the Award for Research Achievement of the Year to **Jure Leskovec**. The “Information Lemon” for the least appropriate information service went to the long-term care information system. The “Information Strawberry” for the best system or service in 2025/2026 went to the electronic identity card. Our warmest congratulations to all the award winners.

Our vision remains clear: to recognize, use, and shape the opportunities of digital transformation, and to create an information society that benefits all of its members. We thank all participants for their contributions. We look forward to shaping, together, achievements that will determine the fate of Slovenia and of the world.

Mojca Ciglarič, Chair of the Program Committee
Matjaž Gams, Chair of the Organizing Committee

KONFERENČNI ODBORI

CONFERENCE COMMITTEES

International Programme Committee

Vladimir Bajic, South Africa
Heiner Benking, Germany
Se Woo Cheon, South Korea
Howie Firth, UK
Olga Fomichova, Russia
Vladimir Fomichov, Russia
Vesna Hljuz Dobric, Croatia
Alfred Inselberg, Israel
Jay Liebowitz, USA
Huan Liu, Singapore
Henz Martin, Germany
Marcin Paprzycki, USA
Claude Sammut, Australia
Jiri Wiedermann, Czech Republic
Xindong Wu, USA
Yiming Ye, USA
Ning Zhong, USA
Wray Buntine, Australia
Bezalel Gavish, USA
Gal A. Kaminka, Israel
Mike Bain, Australia
Michela Milano, Italy
Derong Liu, Chicago, USA
Toby Walsh, Australia
Sergio Campos-Cordobes, Spain
Shabnam Farahmand, Finland
Sergio Crovella, Italy

Organizing Committee

Matjaž Gams, chair
Mitja Luštrek
Lana Zemljak
Vesna Koricki
Mitja Lasič
Blaž Mahnič

Programme Committee

Mojca Ciglarich, chair
Bojan Orel,
Franc Solina,
Viljan Mahnič,
Cene Bavec,
Tomaž Kalin,
Jozsef Györkös,
Tadej Bajd
Jaroslav Berce
Mojca Bernik
Marko Bohanec
Ivan Bratko
Andrej Brodnik
Dušan Caf
Saša Divjak
Tomaž Erjavec
Bogdan Filipič
Andrej Gams
Matjaž Gams
Mitja Luštrek
Marko Grobelnik
Nikola Guid

Marjan Heričko
Borka Jerman Blažič Džonova
Gorazd Kandus
Urban Kordeš
Marjan Krisper
Andrej Kuščer
Jadran Lenarčič
Borut Likar
Janez Malačič
Olga Markič
Dunja Mladenich
Franc Novak
Vladislav Rajkovič
Grega Repovš
Ivan Rozman
Niko Schlamberger
Gašper Slapničar
Stanko Strmčnik
Jurij Šilc
Jurij Tasič
Denis Trček
Andrej Ule

Boštjan Vilfan
Baldomir Zajc
Blaž Zupan
Boris Žemva
Leon Žlajpah
Niko Zimic
Rok Piltaver
Toma Strle
Tine Kolenik
Franci Pivec
Uroš Rajkovič
Borut Batagelj
Tomaž Ogrin
Bojan Blažica
Matjaž Kljun
Robert Blatnik
Erik Dovgan
Špela Stres
Anton Gradišek
Borut Trpin
Anka Slana Ozimič

KAZALO / TABLE OF CONTENTS

<i>Konferenca o umetni inteligenci za pospeševanje doseganja ciljev trajnostnega razvoja / Conference on AI Accelerating the Sustainable Development Goals (AI4SDG)</i>	1
PREDGOVOR / FOREWORD	3
ClimateRAG: AI-Assisted Comparative Analysis of National Climate Commitments and Open Corpus Coverage / Saha Rohan, Nyalala Innocent	5
Revisiting Sustainability by Design in AI Protocol Governance: An Empirical Review of Comparative DAO and Corporate-Led Standards for the SDGs / Luyao Zhang, Wang Yutian	9
Mapping Artificial Intelligence Research for Climate Change Resilience, Disaster Risk Reduction, and Environmental Sustainability: A Bibliometric Perspective on Accelerating the SDGs / Parvathy S. U., Kani Kolil Vysakh, R. Guruvayur Sivaramakrishnan, Achuthan Krishnashree	13
The Emerging AI Power Crisis or Collusion of Two Revolutions / Ghedhioui Imen, Shaikh Iqra, Meyerhoff Nielsen Morten, Martins João	17
Decolonising the Digital Classroom: A Theoretical Case for AI-Mediated Indigenous Knowledge Integration in Open Distance eLearning / Age Terungwa James	21
AI-Enabled Knowledge Infrastructures for Social Innovation: Business Value, Adoption, and Governance / Ali Abbas Syed Muhammad, Gams Matjaž	25
AIDE: Leadership in AI Literacy for Inclusive SDG-Aligned Learning in Marginalized Educational Communities / Mohd Yusuf Mohammad Rizal, Jaris Julieyana	29
Responsible AI Implementation in Education: Protecting Children's Rights in Educational Chatbot Interactions with Younger Learners / Yeo Bernice	33
Governing AI in Primary Education: A Process-Aware Policy Framework / Geoffre Trystan, Geoffre Thierry, Gulati Asheesh	36
Artificial Intelligence in Inclusive Classrooms: Exploring Teachers' Perceptions, Challenges, Opportunities and Professional Development Needs for Children with Disabilities / Bansal Sneha, Jogan Sushma N., Badiger Prakash K.	40
Rethinking 'Good' in AI for Good: Communities as Judges of University-Led AI for Sustainable Development / Bong Yang, Sailaja Neelima, Chen Minjia, Annamali Devisakti, Tada Jin, Cellérier Leïla, Iqbal Muhammad Zahid, Pitogo Vicente	44
Artificial Intelligence Readiness in Brazilian Schools: A Nationwide Survey of Teachers and Educational Challenges / Florentino Bruno Rafael, Sestito Camila, Carvalho André Carlos Ponce de Leon Ferreira, Cruz Wellington, Bonidia Robson Parmezan	48
Beyond AI Governance: Digital Space Resilience and the Bad Actor Problem / Hallensleben Sebastian, Bosco Francesca	52
AI-Based Analysis of Online Data for Climate Hazard Resilience: From Digital Traces to SDG13 Indicators / Erokhin Dmitry, Komendantova Nadejda	56
From Documents to Signals: Toward an AI-powered Policy Intelligence Framework / Oda Masato, Kobayashi Masamichi, Oda Kazue, Fuse Tomoyuki, Taya Sho	60
Documented, Not Governed: Public Algorithm Registers and the Runtime Gap in EU AI Act Deployer Obligations / Nasteski Zoran, Jankulovski Hristijan, Bogoeski Teodor	64
Sustainable Artificial Intelligence Infrastructure and Digital Sovereignty: A Proposal for a Sustainability Impact Assessment and a Data Center Responsibility Index / Quintiliano Leonardo David	68
Equitable Urban Flood Risk Mapping with Ensemble Machine Learning and Google Earth Engine: A Responsible AI Approach for Climate-Resilient Cities in Data-Scarce Regions / Idowu Peter Damilare	72
Generative AI Readiness for Future Skills and Lifelong Learning: Engineering and Non-Engineering Students at UTB and Universiti Malaya / Yunus Sharina	75
AI-Enabled Collaborative Urban Design for Sustainable Development: A Machine Learning Approach to Participatory Planning in Isfahan / Ezzatian Shohreh	79
Federated Causal Inference for Privacy-Preserving Urban Health Policy Evaluation: A Smart City Proof-of-Concept / Rajeev Devi, Guruvayur Sivaramakrishnan R.	83

Artificial intelligence, workforce capabilities and sustainability in tourism distribution / Sanoussi Bendjedid Rachad, Mang Stefan, Maier Janine, Peña-Ramírez Camilo, Benjelloun Andaloussi Omar, Marhraoui Mohamed Amine.....	87
<i>Indeks avtorjev / Author index</i>.....	91

Zbornik 29. mednarodne multikonference
INFORMACIJSKA DRUŽBA – IS 2026
Zvezek M

Proceedings of the 29th International Multiconference
INFORMATION SOCIETY – IS 2026
Volume M

Konferenca o umetniški inteligenci in pospeševanje
doseganja ciljev trajnostnega razvoja
Conference on Artificial Intelligence and Accelerating the Sustainable
Development Goals (AI4SDG)

Uredniki / Editors

Joao Pita Costa, Joao Paulo Vitor, Anja Polajner, Andrej Srakar

<https://is.ijs.si>

9. oktober 2026 / 9 October 2026
Ljubljana, Slovenia

FOREWORD

This workshop is colocated with the 29th International Multiconference IS2026 held on October 5-9 2026 in Ljubljana, Slovenia. It is hosted at the Jozef Stefan Institute, organized by IRCAI in cooperation with ICAIRE and CIAAM at the University of Sao Paulo (Brazil), supported by the NAIXUS excellence Network for Artificial Intelligence, Knowledge and SUsustainable development, and by ELIAS the European Lighthouse of AI for Sustainability.

The “AI Accelerating Sustainable Development Goals (AI4SDG)” workshop, hosted by Naixus, aims to harness Artificial Intelligence (AI) for advancing Sustainable Development Goals (SDGs) with a focus on how AI can support inclusive, ethical, human-centered, and sustainable development across sectors and regions. The workshop seeks the needed discussion of experts and innovators on interdisciplinary research addressing both technological innovation and societal impact, emphasizing practical applications, governance frameworks, policy implications, and responsible deployment of AI for sustainability. This event features a Half-Day Hybrid Workshop, from 9AM to 12PM CEST followed by an online event from 2PM to 4PM CEST. Following the success of past events, such as the AI in Africa & SDGs workshop in Malmo in 2024, it is designed to stimulate discussion, encourage collaboration, and showcase AI solutions, the workshop seeks to connect researchers, practitioners, policymakers, and communities worldwide.

By showcasing impactful AI initiatives and facilitating the exchange of knowledge and experiences, this initiative aims to strengthen collaboration between AI and SDG communities worldwide. It seeks to identify practical approaches for applying AI to climate action and to support progress towards the Sustainable Development Goals through innovative technological solutions. Bringing together participants from diverse backgrounds, the workshop provides a platform for advancing sustainable development by exploring how AI can contribute to addressing global environmental and societal challenges.

ClimateRAG: AI-Assisted Comparative Analysis of National Climate Commitments and Open Corpus Coverage

Rohan Saha

School of Engineering and Science
IIT Madras Zanzibar
Zanzibar, Tanzania
zda24b009@iitmz.ac.in

Innocent Nyalala

School of Engineering and Science
IIT Madras Zanzibar
Zanzibar, Tanzania
innocent@iitmz.ac.in

Abstract—Climate governance increasingly depends on rapid access to policy texts, including Nationally Determined Contributions (NDCs), national climate laws, and scientific assessments, yet these sources are dispersed, heterogeneous, and difficult to compare across jurisdictions. Keyword search misses policy intent and cross-document patterns, while general-purpose large language models can generate plausible but unverifiable claims. For high-stakes climate accountability and public participation, AI systems must be not only useful but also transparent, traceable, and auditable. We introduce *ClimateRAG*, a provenance-first Retrieval-Augmented Generation (RAG) pipeline that treats relational structure as the source of truth. ClimateRAG ingests the ClimatePolicyRadar open corpus (CC-BY 4.0) and stores 12,498 passage-level chunks from 78 documents across 55 countries in a normalized PostgreSQL schema linking countries, documents, chunks, embeddings, and queries. Each passage is encoded with the all-MiniLM-L6-v2 sentence-transformer (384 dimensions) and retrieved using an HNSW index via pgvector, enabling sub-second semantic search without an external vector store. At inference time, a conservative country-detection pre-filter optionally constrains retrieval to the relevant jurisdiction; retrieved passages are then provided to a LLaMA 3.1 model for answer synthesis. Every answer is returned together with its originating country, document title, publication year, and exact supporting passage, enabling verification and downstream audit. Applying ClimateRAG to the current corpus surfaces policy-relevant comparative signals and dataset-level blind spots. In the current indexed subset, we observe indicative patterns such as a divide between quantified, time-bound mitigation targets and narrative sectoral priorities without binding figures; regionally recurring adaptation themes (e.g., food security and water stress in Small Island Developing States and Sub-Saharan Africa versus infrastructure resilience in higher-income settings); and frequent conditionality on international climate finance among lower-income NDCs. We also identify a structural coverage gap in the open corpus itself: 38 of 55 countries are represented by a single document, African nations account for under 20% of documents, and major emitters (including China and the United States) are sparsely covered,

constraining the depth of comparative analysis. ClimateRAG is deployed as an interactive Streamlit application for policy querying, pipeline inspection, and schema exploration. By pairing semantic retrieval with explicit provenance and logging, the system supports trustworthy, human-centered AI for SDG 13 (Climate Action) and capacity-building for SDG 4 (Quality Education), while providing a data-driven basis for prioritizing future corpus expansion and responsible AI governance in climate decision-making.

Index Terms—retrieval-augmented generation, climate policy, NDCs, provenance, open data, corpus coverage, SDG 13, SDG 4, large language models

I. INTRODUCTION

Comparing national climate commitments is difficult because key sources are spread across many document types and are not easy to search in a consistent way. Simple keyword searches often miss meaning and context, and general-purpose language models can produce fluent answers that are difficult to verify against official text. For climate accountability, users need tools that support comparison while maintaining a clear link from each claim to the source.

This need is especially clear for SDG 13 (Climate Action), where policymakers, journalists, educators, and civil society groups often work with limited time and access to specialist tools. Transparent provenance is also important for SDG 4 (Quality Education) because it supports learning based on primary sources rather than unverified summaries.

This paper presents ClimateRAG, a retrieval-augmented generation (RAG) system for question answering over climate policy documents with explicit provenance. ClimateRAG stores documents and passage-level text in a normalized PostgreSQL schema and links each retrieved passage to its country and parent document through foreign keys. It ingests the ClimatePolicyRadar open corpus [6] (CC-BY 4.0), which includes NDCs, national climate laws, and IPCC/IPBES material, and exposes the pipeline through an interactive Streamlit interface.

Our contributions are as follows: (i) a relational database-first RAG design that preserves end-to-end provenance from answers to passages; (ii) an end-to-end prototype over an

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Information Society 2026, 5–9 October 2026, Ljubljana, Slovenia

© 2026 Copyright held by the owner/author(s).

DOI: 10.70314/is.2026.ai4sdg.25

open climate policy corpus [6]; and (iii) a corpus coverage analysis that reports where open climate documentation is sparse, which is relevant for SDG 13 stakeholders. The remainder of this paper describes the system design, discusses corpus coverage and example analyses, and outlines planned evaluation and future work.

II. RELATED WORK

Retrieval-augmented generation (RAG) combines passage retrieval with language model generation to ground answers in evidence [1]. Dense retrieval methods such as DPR [2] and retrieval+generation pipelines for open-domain QA [3] provide core building blocks for modern RAG systems. For high-stakes use, retrieval quality and careful evaluation are essential [7]. Many systems rely on approximate nearest-neighbor (ANN) indexes such as HNSW [4], which is available in PostgreSQL through pgvector [5].

Policy-focused systems, such as PolicyBot [8] show that multistage retrieval, conservative generation, and empirical evaluation can improve reliability for policy question answering. Clinical guideline RAG pipelines also provide a template for end-to-end evaluation of retrieval and citation grounding in high-stakes domains [9]. ClimateRAG complements these efforts by focusing on a relational database-first provenance model and by surfacing corpus coverage gaps as a practical constraint on the analyses that are supported.

III. SYSTEM ARCHITECTURE

ClimateRAG is organized into 3 layers: a relational data layer, a retrieval layer, and an application layer. Fig. 1 illustrates the end-to-end pipeline.

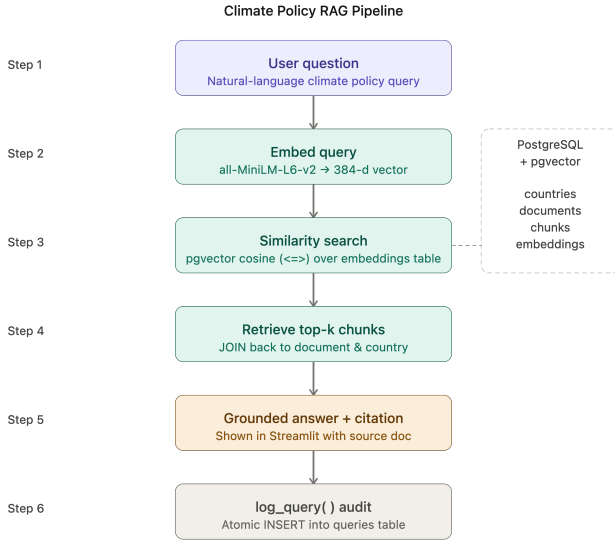


Fig. 1. End-to-end ClimateRAG pipeline architecture.

A. Data Layer: Relational Schema

The data layer is a PostgreSQL 16 database organized into 5 interlinked tables (Third Normal Form, 3NF): `countries`, `documents`, `chunks`, `embeddings`, and `queries`. The

schema follows a simple hierarchy: each country can have many documents, each document is split into many text chunks, and each chunk has 1 embedding vector. The tables are normalized so that core metadata is stored once and can be joined back when needed.

Country metadata (`region`, `continent`) is stored exclusively in the `countries` dimension table, eliminating the transitive dependency that would arise if geographic attributes were stored directly on documents. Referential integrity is enforced throughout: `ON DELETE CASCADE` propagates deletions from documents to chunks and embeddings, while `ON DELETE SET NULL` preserves query audit records and document rows if a parent country is removed. A `CHECK` constraint on `documents.doc_type` restricts values to NDC, Law, IPCC, and IPBES, enforcing document-type classification at the database level.

B. Data Ingestion and Corpus Preparation

The primary data source is the ClimatePolicyRadar open corpus [6], hosted on Hugging Face and released under a CC-BY 4.0 license. The corpus provides pre-chunked passages from 3 families of documents: IPCC AR6 assessment chapters, NDCs submitted under the Paris Agreement, and national climate laws (plus IPBES material).

Raw records were filtered and reshaped into the 5-table schema through a 4-stage ingestion pipeline. Country names were mapped to ISO 3166-1 alpha-2 codes to populate the `countries` dimension table and eliminate inconsistent free-text naming across source documents. Each document was classified by type based on corpus metadata and enforced via the `doc_type` `CHECK` constraint.

Because the ClimatePolicyRadar dataset is distributed with passage-level text segments, ClimateRAG does not introduce a new chunking algorithm in this version. Instead, we apply simple quality controls before indexing. We remove empty or near-empty segments, normalize whitespace, and filter obvious boilerplate (for example, copyright headers and repeated navigation text) using lightweight rule-based heuristics. We do not currently apply systematic deduplication across documents; near-duplicate removal is treated as an implementation-dependent step and is reported as a limitation for reproducibility (Sec. VI).

Finally, every retained chunk was passed through the `sentence-transformers/all-MiniLM-L6-v2` model to produce a 384-dimensional dense embedding. The final corpus comprises 12,498 passage-level chunks drawn from 78 documents across 55 countries.

C. Retrieval Layer: Semantic Search with Provenance

Each chunk embedding is stored in the `embeddings` table as a native `vector(384)` column via the `pgvector` extension [5], indexed using a Hierarchical Navigable Small World (HNSW) structure with cosine similarity (`vector_cosine_ops`), configured with $m = 16$ and $ef_construction = 64$. This enables sub-second approximate

nearest-neighbor retrieval that scales with corpus growth without requiring a dedicated external vector store.

At inference time, the user’s natural-language query is embedded using the same `all-MiniLM-L6-v2` model, ensuring that the query vector occupies the same embedding space as the stored chunk vectors.

A country-detection pre-filter can optionally constrain semantic search to a jurisdiction when the question is clearly country-scoped. In the current prototype, the filter is a conservative rule-based matcher over country names and common aliases stored in the `countries` table. If multiple countries match, the system does not apply the pre-filter and instead falls back to corpus-wide retrieval. If no country is detected, retrieval also falls back to corpus-wide search. This design prioritizes recall over aggressive scoping, but it can still fail when a query uses uncommon aliases, refers to regions rather than countries, or asks for multi-country comparison.

The top- k most similar passages are retrieved via a `pgvector` cosine similarity query joined back to `chunks` and `documents` for full metadata provenance. Retrieved passages, together with their document title, country, and publication year, are passed to a LLaMA 3.1 language model (via Groq API) for grounded answer synthesis. To reduce hallucinations, the generation prompt instructs the model to answer only using the provided passages and to return an explicit “insufficient evidence” response when the context does not support a claim. Every generated answer is accompanied by the exact retrieved passage(s) from which it was derived, making the system’s reasoning inspectable.

D. Audit and Query Logging

Every user interaction is persisted to the `queries` table through the `log_query` stored procedure, which performs an atomic `INSERT` recording the question text, generated answer, and the primary key of the top retrieved chunk. This design prevents partial writes if the application crashes mid-operation and ensures the logged timestamp reflects server time rather than a potentially misconfigured client clock.

We emphasize that this is a minimal audit record sufficient to support basic provenance tracing in the prototype. For stronger auditability and reproducibility, the log should be extended to store the full top- k retrieval set (chunk IDs and similarity scores), the country-scope decision (pre-filter on/off and matched country codes), the prompt template and decoding parameters, and explicit embedding and LLM model identifiers and versions. When possible, logging should avoid storing sensitive user data; in deployments that accept real user queries, logs should be access-controlled and retention-limited.

E. Application Layer: Streamlit Interface

The application layer is a multi-page Streamlit web application providing 5 dedicated views: a live policy query interface, a RAG pipeline inspector, a schema explorer with live SQL execution, a corpus statistics dashboard displaying the 55-country distribution, and a design justification page explaining key architectural decisions. The query interface

displays each answer alongside the source document title, country, publication year, and retrieved passage, so users can verify every claim against its origin rather than trusting an unconstrained generation.

IV. EVALUATION AND PRELIMINARY RESULTS

Table I presents preliminary empirical results from the deployed system over the indexed corpus.

TABLE I
PRELIMINARY SYSTEM EVALUATION RESULTS.

Metric	Value
Vector retrieval latency (HNSW, top-5, warm cache)	0.69 ms
Vector retrieval latency (cold cache, first query)	101.8 ms
Country pre-filter precision/recall (explicit queries, $n=11$)	100%/100%
Manual citation support rate	70% (7/10)

Vector retrieval latency was measured using EXPLAIN ANALYZE over the `pgvector` HNSW index on a warm cache, confirming sub-millisecond approximate nearest-neighbor search over 12,498 embeddings. Citation support was assessed by manually evaluating 10 queries against countries present in the corpus: an answer was marked as supported if every key claim was traceable to the retrieved source passage. Three unsupported cases reflect passages where the correct document was retrieved but the specific answer was absent from the indexed chunk, rather than system hallucination. These results are preliminary and intended to characterize prototype behavior rather than benchmark against baselines. A full quantitative evaluation including Recall@ k , MRR, and citation-faithfulness metrics (e.g., RAGAS) is planned as corpus coverage expands.

A. Reproducibility Notes

To support replication of the current prototype without introducing new experiments, we summarize key fixed configuration choices used in this paper: PostgreSQL 16 with `pgvector` HNSW (cosine distance; $m = 16$, `ef_construction = 64`), a `vector(384)` embedding column, `sentence-transformers/all-MiniLM-L6-v2` embeddings, top- $k = 5$ retrieval, and a generation prompt that requires answers to be grounded only in retrieved passages with an “insufficient evidence” fallback. We recommend reporting exact library versions, the dataset snapshot ID, and LLM decoding parameters (temperature, max tokens) in an artifact release.

V. CORPUS ANALYSIS (CASE STUDIES)

ClimateRAG can support exploratory, provenance-grounded case studies of how national commitments are expressed in the currently indexed subset of the ClimatePolicyRadar corpus. We report corpus statistics and illustrative analyses, but avoid generalizing beyond what the corpus coverage supports.

A. Corpus Coverage and Geographic Distribution

Table II summarizes the final corpus statistics. Of the 55 countries represented, 38 appear with exactly 1 document. This reflects the distribution of the underlying open corpus rather than a sampling decision. Japan and India lead coverage with 5 documents each, followed by Argentina with 3 documents. 14 countries appear with 2 documents, while the remaining 38 are represented by a single passage source.

TABLE II
CORPUS COVERAGE SUMMARY.

Property	Value
Countries represented	55
Total documents	78
Total text chunks	12,498
Countries with exactly 1 document	38 (69%)
Countries with ≥ 2 documents	17 (31%)
Share of documents from African nations	<20%
Top covered countries	Japan, India (5 each)

Geographic representation is uneven in ways that carry policy significance. African nations account for fewer than 20% of total documents in the corpus despite representing the continent most acutely exposed to climate risk. Major emitters including China and the United States are present but sparsely covered (China with 1 document and the United States with 2). This limits the depth of cross-national comparison on mitigation ambition that the system can currently support. These gaps are not artifacts of the ClimateRAG ingestion process; they reflect the structure of the underlying open corpus and are themselves a finding of significance for the open data community and for organizations such as ClimatePolicyRadar working to expand documentation coverage.

B. Mitigation Target Specificity: A Comparative Dimension

Within NDC-tagged documents in the current corpus, countries differ in how explicitly they state mitigation targets. Where multiple submissions are available (e.g., Japan and India), limited within-country comparison suggests that later submissions more often include quantified, time-bound targets, while earlier submissions more often describe sectoral priorities without binding figures.

C. Geographic Coverage as a Policy Finding

Table II shows that the indexed open corpus is unevenly distributed across countries and regions (e.g., many countries with a single document and limited coverage of major emitters). These coverage constraints bound what cross-national comparisons the system can support and provide a concrete basis for prioritizing future corpus expansion.

VI. LIMITATIONS AND REPRODUCIBILITY

ClimateRAG is a working system but not yet a fully benchmarked research artifact. Key limitations are: (i) the current indexed subset covers 78 documents across 55 countries,

which bounds the comparisons supported; (ii) retrieval is currently dense-only, without hybrid retrieval or re-ranking baselines; (iii) the country pre-filter is rule-based and evaluated only on explicit single-country queries ($n = 11$), while alias and multi-country behavior remains unevaluated; and (iv) the audit log is minimal and should be extended to store top- k retrieval results, similarity scores, and explicit model, prompt, and decoding identifiers.

We also do not evaluate downstream decision impact. ClimateRAG is meant to support users in reading and comparing source text, not to replace expert judgment in policy analysis.

To support reproducibility, future work will include releasing the database DDL, ingestion scripts, and evaluation queries, along with a small human-verified QA set for retrieval and citation-faithfulness evaluation.

VII. CONCLUSION AND FUTURE WORK

This paper presented ClimateRAG, a provenance-aware retrieval-augmented generation pipeline for climate policy question answering. The system combines a normalized PostgreSQL schema with pgvector semantic search and LLaMA 3.1 answer synthesis so that users can trace each answer back to specific source passages. Queries and outputs are logged to support basic audit and later analysis. Using the current indexed subset of the ClimatePolicyRadar corpus, we also report coverage constraints that affect what comparisons are possible (Table II). These gaps limit cross-country analysis and motivate targeted corpus expansion toward broader coverage of the UNFCCC parties.

Future work will expand the corpus, add hybrid sparse-dense retrieval, e.g., BM25 + embeddings with reranking, and complete the quantitative evaluation described in Section IV.

ACKNOWLEDGMENT

The authors thank the ClimatePolicyRadar project for making the climate policy corpus available under a CC-BY 4.0 license, and the IRCAI NAIXUS network for the opportunity to present this work at the AI4SDG Workshop 2026.

REFERENCES

- [1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. K  ttler, M. Lewis, W. Yih, T. Rockt  schel, and others, “Retrieval-augmented generation for knowledge-intensive NLP tasks,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.
- [2] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih, “Dense passage retrieval for open-domain question answering,” in *Proc. 2020 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, 2020.
- [3] G. Izacard and E. Grave, “Leveraging passage retrieval with generative models for open domain question answering,” in *Proc. 16th Conf. European Chapter of the Association for Computational Linguistics (EACL)*, pp. 874–880, 2021.
- [4] Y. A. Malkov and D. A. Yashunin, “Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 824–836, 2018.
- [5] A. Kane and others, “pgvector: Open-source vector similarity search for Postgres,” GitHub, 2021. [Online]. Available: <https://github.com/pgvector/pgvector>.
- [6] ClimatePolicyRadar, “All document text data,” Hugging Face dataset, CC-BY 4.0, 2023. [Online]. Available: <https://huggingface.co/datasets/ClimatePolicyRadar/all-document-text-data>.

Revisiting Sustainability by Design in AI Protocol Governance: An Empirical Review of Comparative Open and Corporate-Led Standards for the SDGs*

Yutian Wang
Duke Kunshan University
Suzhou, China

Luyao Zhang[†]
Duke Kunshan University
Suzhou, China

Abstract

As artificial intelligence (AI) agents enter production infrastructure, the protocols that let them interoperate also shape how that infrastructure is governed and sustained. This paper revisits evidence from our earlier comparative governance study through a Sustainability by Design (SbD) lens. That empirical study compares two standards for AI-agent interoperability: Ethereum Request for Comments 8004 (ERC-8004), developed through an open Ethereum standards process, and Google’s Agent2Agent (A2A), governed by a corporate consortium. It uses a large language model (LLM)-powered pipeline that combines automated annotation, neural topic modeling, and multi-layer network analysis to examine socio-technical power structures across large public governance repositories. The study reports two contrasting governance architectures. ERC-8004 uses open participation, public review, and decoupled deployment, whereas A2A assigns binding authority to an eight-seat Technology Steering Committee. It also finds distinct innovation infrastructures. The ERC community concentrates on constitutive questions about trust and security, including what to build and why, while the consortium spreads attention across executive engineering questions about how to implement, document, and deliver the protocol. Both communities show high participation inequality, and corporate contributors span roughly twice as many themes as ERC contributors. This review asks: (1) How do these governance architectures produce distinct SDG-relevant signatures? (2) What design principles do those signatures suggest for sustainable AI governance? We interpret the reported institutional, discursive, and network patterns in relation to SDGs 8, 9, 10, 11, 12, 16, and 17. The analysis identifies capacities for transparency, participation, contestability, and cross-protocol coordination. It does not treat

those capacities as evidence of achieved SDG impact. We argue that sustainable AI infrastructure needs a corrective feedback loop between designed charters and governance in practice, an institutional commitment associated with SDG 16 (Peace, Justice, and Strong Institutions). The review brings together the prior study’s computational evidence, organizational research, and sustainable development to derive design principles for sustainable AI governance intended for scholars, policymakers, and protocol designers.

Keywords

AI protocol governance; Sustainability by Design; open standards governance; interoperability standards; Sustainable Development Goals

1 Introduction

Governance for the Sustainable Development Goals (SDGs) includes the institutions behind AI infrastructure. When autonomous agents negotiate tasks across organizational boundaries, interoperability protocols distribute authority, innovation work, and participation. Research on AI for the SDGs more often evaluates applications than the protocols beneath them.

We address this gap by reviewing our earlier comparative governance study [1] through a *Sustainability by Design* (SbD) lens. In this review, an *SDG governance signature* is an observable institutional, discursive, or participation pattern that bears on an SDG-relevant capacity. A signature does not by itself demonstrate SDG achievement. Figure 1 summarizes the paper’s central analytical claim: governance design sets rules for access, authority, and transparency, and public records then show what communities discuss, who participates, and how participants connect. We use these signatures to consider implications for SDGs 8, 9, 10, 11, 12, 16, and 17.

The earlier study compares two protocols that address the same AI-agent interoperability problem and have public governance records [2]. The data and analysis materials are available at <https://github.com/Global-Nomad-Nexus/erc8004-a2a-case-study>. Ethereum Request for Comments 8004 (ERC-8004), developed through an open Ethereum standards process, is an open Ethereum standard that enables agents to discover, choose, and interact across organizational boundaries without pre-existing trust [3].

*An earlier version of this paper was presented at the 8th International Conference on Management and Organization (ICMO 2026), held in Pržno, Montenegro, from 3 to 6 June 2026. The conference, themed “Sustainability by Design – Rethinking Strategy, People & Digital Futures,” was hosted by the Montenegrin Academy of Management (MAOM), in collaboration with the Slovenian Academy of Management, the Croatian Academy of Management, and the Faculty of Economics Podgorica at the University of Montenegro. We are grateful to the conference organizers and participants for their valuable comments and constructive feedback, which helped improve this paper.

[†]Corresponding author: Luyao Zhang (lz183@duke.edu), Digital Innovation Research Center and Social Science Division, Duke Kunshan University. Address: Duke Avenue No.8, Kunshan, Suzhou, Jiangsu, China, 215 316.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Information Society 2026, 5–9 October 2026, Ljubljana, Slovenia

© 2025 Copyright held by the owner/author(s).

<https://doi.org/10.70314/is.2026.ai4sdg.1>

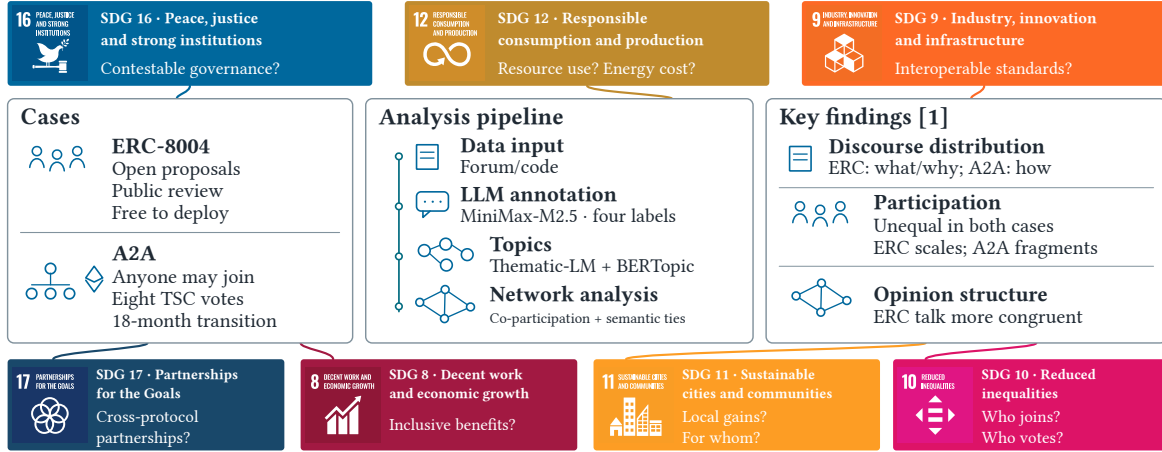


Figure 1: White panels show the earlier comparison [1], with governance details checked against [3, 4, 7, 8]. Color panels pose SDG questions, not measured outcomes or protocol energy footprints. SDG icons: United Nations (<https://sdgs.un.org/goals>); use does not imply UN endorsement.

Its proposal and review process is public, while deployment is separate from standards discussion [3, 4]. Google’s Agent2Agent (A2A) is an open protocol for communication and interoperability between opaque agentic applications. Google launched it in April 2025 and donated it to the Linux Foundation in June 2025 [5, 6]. Its charter permits anyone to contribute or become a maintainer, while an eight-seat Technology Steering Committee (TSC) holds initial voting authority. The 18-month provision concerns future TSC composition [7, 8]. This design holds the technical domain broadly constant while comparing open entry in both cases with different routes to binding decisions [7, 9].

The review addresses two questions:

- **RQ1:** How do contrasting AI protocol governance architectures produce distinct SDG-relevant signatures?
- **RQ2:** What design principles do these signatures imply for sustainable AI governance?

We reinterpret this prior empirical comparison through SbD, develop a signature framework covering seven goals, and propose design principles for sustainable AI governance.

2 Reviewed evidence and analytical lens

The earlier empirical study reports three types of evidence [1]: decision architectures that document entry and binding authority, discourse distributions that distinguish constitutive design from implementation, and participation networks that measure concentration and cross-standard coordination. The original study contains the methodological details, validation procedures, and robustness checks.

This review does not repeat that empirical analysis. It interprets the reported evidence as SDG-relevant governance capacities, considers the design corrections those capacities may require, and keeps institutional capacity distinct from achieved societal impact.

3 Findings

3.1 Governance Architecture Signatures

The prior study reports contrasting governance architectures. ERC-8004 has a public proposal and review process and separates deployment from standards discussion. A2A reserves binding authority for its eight-seat TSC [7]. EIP-1’s rough-consensus provision concerns Core EIPs [4]; A2A’s TSC makes binding decisions under its charter and GitVote procedure [7, 8].

Figure 2 adds this review’s interpretation. Its left annotation presents a *Design* → *Emerge* → *Reality* → *Correct* loop. The loop compares authority in practice with the stated rules, then uses that comparison to revise the governance design.

3.2 Different Objects of Deliberation

The prior study finds that A2A devotes nearly twice the share of its discourse to Process arguments as ERC-8004 (25.4% vs. 13.9%; $p < .001$). Within ERC-8004, Technical arguments dominate the design stage, while Process arguments rise during ratification ($p < .001$). ERC-8004 participants focus on what should define the protocol, especially trust mechanisms, security, identity, and specification. A2A participants spread their attention across implementing, documenting, integrating, and shipping an established protocol.

The prior study’s thematic analysis shows that ERC-8004 centers on Trust & Security and Protocol Specification, whereas A2A covers Documentation, software development kits (SDKs), Tooling, and other execution themes. Both communities have high participation inequality. A small core accounts for most interactions, while nearly half of A2A participants and one third of ERC-8004 participants remain isolated. Corporate contributors span roughly twice as many themes as ERC contributors, consistent with specialization by design in A2A and specialization by attrition in ERC-8004 [10].

This review interprets that deliberative split in relation to SDG 9 (Industry, Innovation and Infrastructure). Figure 3 presents the thematic differences reported in the prior study [1] and keeps them separate from this review’s SDG association.

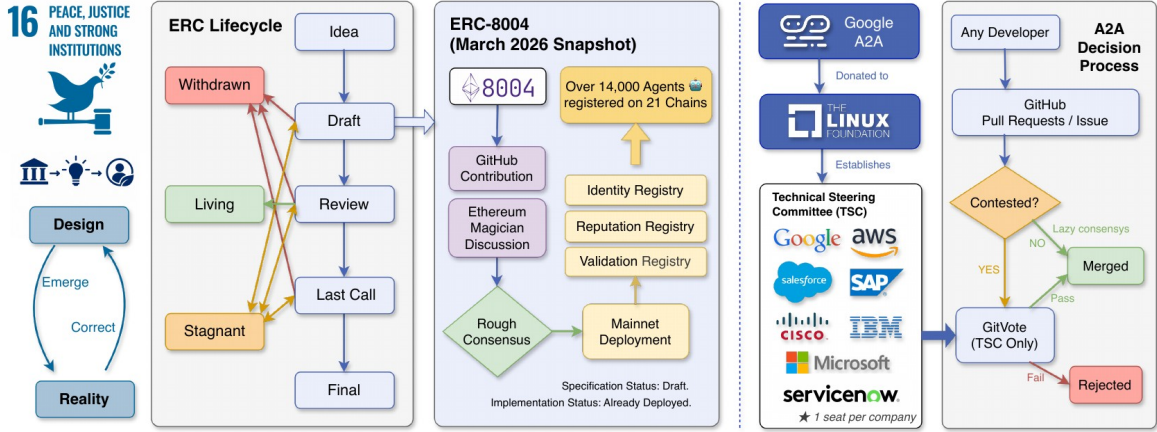


Figure 2: The prior study's contrasting decision architectures [1], with this review's SDG 16 Design → Emerge → Reality → Correct loop.

3.3 Connectivity Depends on the Scale of Observation

The prior study finds different structures at the protocol and ecosystem scales. In the comparison between ERC-8004 alone and A2A, both public networks are fragmented and participation is concentrated. When the ERC corpus expands from 142 ERC-8004 records to 1,664 records across 34 related AI protocols, repeat contributors and cross-standard discussions connect previously separate threads. The ERC giant-component ratio rises from 0.328 to 0.917. A2A's public trace remains fragmented at 0.285 in the expanded analysis [1].

In this review, the difference between scales shows that protocol-level fragmentation can coexist with ecosystem-level coordination. It also means that connectivity cannot be used as a proxy for inclusion.

4 Discussion

The earlier study identifies different governance capacities and risks. ERC-8004 makes constitutive deliberation visible and allows anyone to participate, while A2A coordinates implementation through formal corporate authority. The comparison does not show that open standards advance sustainability or that a corporation obstructs it. We use the SDGs as an assessment lens. They are not evidence that either protocol has achieved societal impact.

4.1 Extended SDG and Policy Implications

Our interpretation extends the institutional comparison beyond SDG 16. ERC-8004 makes questions of trust, security, and identity open to public contestation, while A2A's TSC organizes work across documentation, tooling, and integration. In relation to SDG 9, these patterns suggest two requirements. Open protocols need reliable procedures for resolution and maintenance, and corporate-led protocols need auditable routes for outside challenge. The high isolation reported in both cases is also relevant to SDGs 8, 10, and 11, because access, governance labor, and connection to decisions are distinct outcomes.

The 34-protocol comparison shows why our interpretation must account for scale. Cross-standard contributors form an ecosystem-level capacity for partnership that is relevant to SDG 17, but concentrated participation persists. Connectivity and inclusion therefore require separate measures. Public records offer an analogy to SDG 12 traceability, not proof of sustainable consumption. Environmental footprints need separate measurement: 34 large U.S. Bitcoin mines consumed an estimated 32.3 TWh during August 2022–July 2023 [20]. That proof-of-work result cannot be assigned to ERC-8004 or A2A.

The United Nations Convention on the Law of the Sea (UNCLOS) provides a boundary case for coordination. The law of the sea coordinates states in a cross-border domain through durable institutions and negotiated obligations [11]. AI protocols, by contrast, can fork, recombine, and migrate without territorial continuity. Sustainable AI governance can draw on the durability of international coordination without adopting a state-centered form. Platform regulation [12] and blockchain treasury governance [13] frame authority and accountability; a UNICEF pilot illustrates one funding route for open digital public goods [14].

4.2 UNESCO AI Ethics and Governance Frameworks

UNESCO's 2021 Recommendation on the Ethics of Artificial Intelligence, adopted by 193 Member States, calls for inclusive multi-stakeholder governance, responsibility, auditability, traceability, transparency, non-discrimination, and human oversight [15]. We use these principles to interpret the empirical comparison. ERC-8004's permissionless entry and public deliberation records correspond more closely to participation and auditability. A2A's initial eight corporate TSC voting seats concentrate formal authority despite open contributor and maintainer pathways, creating tension with inclusiveness even as the TSC coordinates implementation. The participation results also show that open entry does not ensure inclusion: both cases remain concentrated.

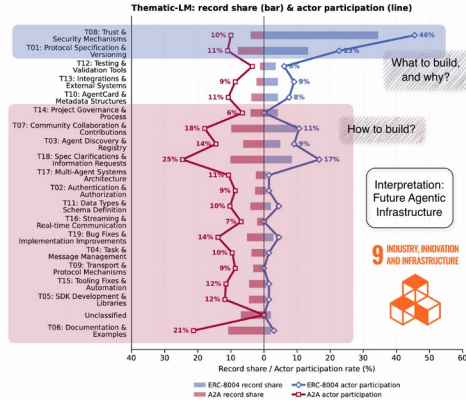


Figure 3: Thematic-LM record share (bars) and actor participation (lines) for what/why and how themes [1]. The SDG 9 marker is this review's interpretation, not measured impact.

UNESCO's Ethical Impact Assessment (EIA) evaluates individual AI systems across their life cycle, whereas its Readiness Assessment Methodology (RAM) evaluates institutional capacity at the country level [16, 17]. Neither tool directly assesses protocol communities. We draw an analogy with EIA and RAM; we do not use either instrument to assign scores. Publicly traceable forum and GitHub records allow external review, contestation, and institutional diagnosis. ERC-8004 provides a broader observable record. A2A's public repository is also observable, but outsiders have less access to binding authority. Under this interpretation, sustainable governance related to SDG 16 requires records that can be assessed and contested, as well as a designed charter.

The *Design* → *Emergence* → *Reality* → *Correct* loop in Figure 2 is a protocol-governance form of UNESCO's monitoring logic. The *Reality* → *Correct* step connects impact monitoring and external review to institutional revision. ERC-8004's public discussion provides a feedback channel, but the protocol still needs explicit procedures for resolution and maintenance. A2A has clearer executive authority, but it needs an auditable route for outsiders to contest and revise binding rules.

Participation inequality further shows the difference between formal access and inclusion in practice. UNESCO's Women4EthicalAI platform seeks equal representation in AI design, use, and deployment [18]. ERC-8004 removes a formal barrier to entry, while A2A concentrates initial TSC voting authority among corporate members. Neither finding demonstrates gender equality because the dataset does not identify gender. Protocol stewards should report representation, isolation, coordination work, and access to binding decisions as separate measures.

At the ecosystem scale, cross-protocol ties create governance capacity relevant to SDGs 9 and 17. Because these ties coexist with concentrated participation, connectivity and inclusion should remain separate governance indicators.

These conclusions are limited by the use of public records, the focus on two governance forms, dependence on natural language processing models [19], and culturally contingent ideas of fairness. Further research should examine other protocol families and whether corrective mechanisms improve

participation and decision quality. Sustainable AI governance needs technical openness and institutions that keep authority observable, connect deliberation across protocol and ecosystem scales, and correct inequalities that those institutions themselves produce.

Acknowledgments

L.Z. acknowledges the support from the National Natural Science Foundation of China (NSFC) under the project titled "Trust Mechanism Design on Blockchain: An Interdisciplinary Approach of Game Theory, Reinforcement Learning, and Human-AI Interactions" (Grant No. 12201266). Y.W. gratefully acknowledges the Summer Research Scholar Program and the Office of Academic Services at Duke Kunshan University for support.

References

- [1] Y. Wang and L. Zhang, "Agentic analysis for agentic infrastructure: An llm-powered pipeline for comparative governance of dao and corporate ai protocols," 2026. [Online]. Available: <https://arxiv.org/abs/2606.26203>
- [2] R. K. Yin, *Case Study Research and Applications: Design and Methods*, 6th ed. SAGE, 2018.
- [3] D. Crapis, "ERC-8004: Trustless agents," Ethereum Magicians Forum, topic 25 098, 2025. [Online]. Available: <https://ethereum-magicians.org/t/erc-8004-trustless-agents/25 098>
- [4] M. Becze, H. Jameson, et al., "EIP-1: EIP purpose and guidelines," 2015. [Online]. Available: <https://eips.ethereum.org/EIPS/eip-1>
- [5] A2A Project, "Agent2Agent (A2A) Protocol," 2025. [Online]. Available: <https://github.com/a2aproject/A2A>
- [6] Google Cloud, "Google Cloud donates A2A to Linux Foundation," 2025. [Online]. Available: <https://developers.googleblog.com/en/google-cloud-donates-a2a-to-linux-foundation/>
- [7] A2A Project, "GOVERNANCE.md," 2025. [Online]. Available: <https://github.com/a2aproject/A2A/blob/main/GOVERNANCE.md>
- [8] a2aproject, "gitvote.yml," GitHub, 2025. [Online]. Available: <https://github.com/a2aproject/A2A/blob/main/.gitvote.yml>
- [9] P. Reineke, R. Katila, and K. M. Eisenhardt, "Decentralization in organizations: A revolution or a mirage?" *Academy of Management Annals*, vol. 19, no. 1, pp. 298–342, 2025.
- [10] L. Dahlander and S. O'Mahony, "Progressing to the center: Coordinating project work," *Organization Science*, vol. 22, no. 4, pp. 961–979, 2011.
- [11] Y. Tanaka, *The International Law of the Sea*, 2nd ed. Cambridge: Cambridge University Press, 2015.
- [12] C. Cennamo, P. Gianiodis, T. Kretschmer, D. McIntyre, P. Ozcan, and M. Van Alstyne, "Platform regulation: Beyond power and size," *Academy of Management Perspectives*, 2025. doi:10.5465/amp.2025.0261.
- [13] D. W. E. Allen, C. Berg, and A. M. Lane, "The blockchain treasury governance dilemma," *Regulation & Governance*, vol. 19, no. 3, pp. 761–776, 2025. doi:10.1111/rego.12659.
- [14] A. Maharajan, "Exploring sustainable funding for digital public goods," UNICEF Office of Innovation, Jan. 7, 2026. [Online]. Available: <https://www.unicef.org/innovation/stories/exploring-sustainable-funding-digital-public-goods>
- [15] UNESCO, "Recommendation on the ethics of artificial intelligence," UNESCO General Conference, 41st session, 2021, accessed: July 19, 2026. [Online]. Available: <https://unesdoc.unesco.org/ark:/48 223/pf0 000 381 137>
- [16] —, "Ethical impact assessment," Global AI Ethics and Governance Observatory, n.d., accessed: July 19, 2026. [Online]. Available: <https://www.unesco.org/ethics-ai/en/eia>
- [17] —, "Readiness assessment methodology: A tool of the recommendation on the ethics of artificial intelligence," 2023, accessed: July 19, 2026. [Online]. Available: <https://unesdoc.unesco.org/ark:/48 223/pf0 000 385 198>
- [18] —, "Women4Ethical AI," n.d., accessed: July 19, 2026. [Online]. Available: <https://www.unesco.org/en/artificial-intelligence/women4ethical-ai>
- [19] T. R. Hannigan, R. F. J. Haans, K. Vakili, H. Tchalian, V. L. Glaser, M. S. Wang, S. Kaplan, and P. D. Jennings, "Topic modeling in management research: Rendering new theory from textual data," *Academy of Management Annals*, vol. 13, no. 2, pp. 586–632, 2019.
- [20] G. Guidi, F. Dominici, N. Steinsult, et al., "The environmental burden of the United States' bitcoin mining boom," *Nature Communications*, vol. 16, art. 2970, 2025. doi:10.1038/s41 467-025-58 287-3.

Mapping AI Research for Climate Change Resilience, Disaster Risk Reduction, and Environmental Sustainability: A Bibliometric Perspective on Accelerating the SDGs

S. U. Parvathy

Center for Cybersecurity Systems
and Networks, Amrita Vishwa
Vidyapeetham, Amritapuri,
Kollam, Kerala, India
parvathysu@am.amrita.edu

Krishnashree Achuthan

Center for Cybersecurity Systems
and Networks, Amrita Vishwa
Vidyapeetham, Amritapuri,
krishna@amrita.edu*

Vysakh Kani Kolil

Center for Cybersecurity Systems
and Networks, Amrita Vishwa
Vidyapeetham, Amritapuri,
Kollam, Kerala, India
vysakh@am.amrita.edu

Sivaramakrishnan R

Guruvayur

Center for Cybersecurity Systems
and Networks, Amrita Vishwa
Vidyapeetham, Amritapuri,
Kollam, Kerala, India
sivaramakrishnangr@am.amrita.edu

u

Abstract

Artificial intelligence (AI) is increasingly used to address environmental challenges and advance the Sustainable Development Goals (SDGs) through predictive analytics, intelligent monitoring, and data-driven decision-making. However, the rapidly expanding literature remains fragmented. This study presents a bibliometric analysis of 9,937 Scopus-indexed publications on AI, climate resilience, disaster risk reduction, and environmental sustainability published between 2020 and 2026. Keyword co-occurrence analysis identified five major themes: climate resilience and environmental risk assessment; intelligent decision support; ecosystem monitoring and restoration; AI-enabled life cycle assessment and resource management; and air quality monitoring and urban sustainability. SDG mapping revealed strong linkages with SDGs 13, 11, 9, 8, and 6. Keywords—Artificial intelligence, Sustainability, SDGs

1 Introduction

Climate change, increasing disaster risks, biodiversity loss, environmental degradation, and the unsustainable use of natural resources represent some of the most pressing global challenges of the twenty-first century [1]. These interconnected issues threaten ecosystems, human well-being, and economic development while hindering progress toward the United Nations Sustainable Development Goals (SDGs) [2]. Despite

global commitments under the 2030 Agenda, recent assessments indicate that many SDG targets remain off track, emphasizing the need for innovative technologies and evidence-based solutions that can accelerate sustainable development. Among these technologies, artificial intelligence (AI) has emerged as a powerful enabler of environmental sustainability by supporting intelligent decision-making, predictive analytics, and resource optimization [3]. Recent advances in AI, including machine learning, deep learning, computer vision, natural language processing, and remote sensing, have enhanced environmental monitoring, climate-risk prediction, disaster preparedness, and resource management. AI is increasingly applied to climate adaptation, disaster forecasting, biodiversity conservation, renewable energy, air quality monitoring, precision agriculture, and sustainable urban planning, contributing to SDGs through resource efficiency, carbon reduction, ecosystem conservation, resilient infrastructure, and environmental governance. Recent bibliometric studies have examined AI, sustainable development, environmental monitoring, and disaster management; however, these studies largely focus on individual domains or broader AI–SDG relationships. A focused mapping of the intersection of climate resilience, disaster risk reduction, and environmental sustainability can therefore provide a complementary understanding of emerging themes and their SDG linkages. The contribution of this study is therefore the application of bibliometric analysis to the integrated examination of AI research at the intersection of climate resilience, disaster risk reduction, and environmental sustainability, together with its thematic and SDG-level structure

Accordingly, this study addresses the following research questions:

1. How has the research landscape evolved in terms of publication trends, leading authors, institutions, journals, and geographical distribution?

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, 5–9 October 2026, Ljubljana, Slovenia
© 2025 Copyright held by the owner/author(s).
<http://doi.org/10.70314/is.2026.ai4sdg.1>

2. What are the intellectual and conceptual structures, and what major themes have emerged in this research domain?
3. What emerging trends, research gaps, and future directions can be identified to advance AI for climate resilience, disaster risk reduction, and environmental sustainability?

2. Methodology

This study adopted a bibliometric approach [4] to thoroughly investigate the state of AI research for environmental sustainability, disaster risk reduction, and climate change adaptation. The Scopus database was used as the primary source. The search strategy employed the following TITLE-ABS-KEY query: TITLE-ABS-KEY(("artificial intelligence" OR "AI" OR "Agentic AI" OR "machine learning" OR "deep learning" OR "neural network*" OR "expert system*" OR "computer vision" OR "natural language processing" OR "generative AI" OR "large language model*") AND ("climate change" OR "climate resilience" OR "climate adaptation" OR "climate mitigation" OR "global warming" OR "environmental sustainability" OR "sustainability" OR "life cycle assessment" or "LCA" or "life cycle thinking" or "LCT" OR "social life cycle assessment" or "S-LCA" OR "life cycle cost assessment" or "LCC" OR "life cycle sustainability assessment" or "LCSA" OR "environmental management" OR "disaster risk reduction" OR "disaster management" OR "disaster resilience" OR "natural hazard*" OR "climate hazard*" OR resilience)) AND PUBYEAR <2019 AND PUBYEAR > 2027 AND (LIMIT-TO (EXACT KEYWORD,"Article") OR LIMIT-TO (EXACT KEYWORD,"Human") OR LIMIT-TO (EXACT KEYWORD,"China") OR LIMIT-TO (EXACT KEYWORD,"Humans") OR LIMIT-TO (EXACT KEYWORD,"Female") OR LIMIT-TO (EXACT KEYWORD,"Male"))AND (LIMIT-TO (DOCTYPE,"ar"))AND (LIMIT-TO (LANGUAGE,"English")) The search was restricted to publications from 2020–2026, English-language articles, and the specified Scopus filters. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework [6] guided the selection process, resulting in 9,937 publications for analysis. Keyword co-occurrence analysis was conducted using VOSviewer to identify the conceptual structure and major thematic clusters [7]. The final dataset was also mapped to the UN SDGs using the Elsevier SDG Mapping Initiative [5].

3. Results

To address RQ1, the research landscape shows rapid and sustained growth in AI research on climate resilience, disaster risk reduction, and environmental sustainability (Fig. 1(A)). Publications increased from approximately 300 in 2020 to 460 in 2021, 670 in 2022, around 1,000 in 2023, nearly 1,850 in 2024, and over 3,000 in 2025. The lower output in 2026 reflects the incomplete publication year rather than a decline. This growth demonstrates the increasing importance of AI in climate, environmental, and disaster-related research. Further evidence of the field's growth and specialization is the increasing use of phrases like machine learning, deep learning, climate change, disaster management, spatiotemporal analysis, forecasting, and explainable AI. [8], [9]. The journal (Fig. 1(D)) and author (Fig. 1 (B)) distributions further highlight the development of the research community. Scientific Reports is the most productive

journal, followed by Science of the Total Environment, Journal of Environmental Management and, PLOS ONE. Ciais, Philippe leads the authors with 17 publications, followed by Zhang, Shihang; Yin, Jiabo; Singh, Vijay P.; Penuelas, Josep J.; Onyelo Wole, Kennedy Chibuzor; Nikoo, Mohammad Reza; and Chen, Xiaohong, with approximately 12 publications each. Their diverse expertise in climate science, hydrology, environmental modelling, remote sensing, and sustainability highlights the multidisciplinary nature of the field. The leading institutions are predominantly Chinese, indicating a strong institutional concentration of research output in China (Fig. 1(E)). Within the retrieved Scopus corpus, China has the highest publication output, followed by the United States and India (Fig. 1(C)). This distribution represents the geographical composition of the retrieved corpus under the applied search and Scopus indexing conditions and should therefore be interpreted descriptively rather than as an unbiased estimate of global research activity.

The field's cognitive and conceptual structure was investigated using topic clustering and term co-occurrence analysis (Fig. 2(H)) in order to answer RQ2. Five main thematic clusters (Fig. 2(F)) emerged from the analysis, indicating that AI research has moved beyond traditional climate prediction to a variety of applications in ecosystem management, intelligent decision support, environmental risk assessment, sustainable resource management, and urban ecological monitoring [10]. Terms like global warming, extreme events, drought, flood control, ground surface temperature, soil organic carbon, remote sensing, spatial distribution, and risk prediction are characteristics of AI-driven climate resilience, prediction, and environmental risk assessment (Cluster 1). This area emphasizes the widespread application of AI through Earth observation and spatial analysis to comprehend climate-related hazards, forecast environmental risks, and enhance climate resilience [11]. AI methods and intelligent decision support for sustainable systems (Cluster 2), includes explainable machine learning, deep learning, neural networks, forecasting, scenario analysis, and spatiotemporal analysis. The prominence of explainable AI indicates increasing attention to transparency and interpretability in environmental decision-making, suggesting a transition from prediction-oriented applications towards AI-supported decision systems [12]. AI for ecosystem monitoring, restoration, and environmental sustainability (Cluster 3), is associated with biodiversity, ecosystems, resilience, sustainability, public health, green innovation, and environmental monitoring. This cluster demonstrates the expansion of AI towards ecosystem assessment, bio-diversity conservation, environmental degradation monitoring, and restoration [13]. AI-enabled life cycle assessment and sustainable resource management (Cluster 4) connects AI with life cycle approaches and resource efficiency. The presence of machine learning, transfer learning, computer vision, classification, IoT, and predictive maintenance indicates growing interest in data-driven environmental performance assessment, resource optimization, and sustainable management [14]. AI for air quality monitoring and urban environmental sustainability (Cluster 5), focuses on air quality prediction, air pollution, black carbon, particulate matter, atmospheric conditions, and urban environments. This theme demonstrates the increasing application of AI, sensor networks, remote sensing, and spatial information for monitoring environmental conditions and supporting sustainable urban



management [15]. To address RQ3, the emerging trends, research gaps, and future directions were examined based on the thematic structure and SDG distribution of the research landscape. The thematic structure indicates increasing diversification of AI applications across climate-risk assessment, decision support, ecosystem monitoring, life cycle-based sustainability assessment, and urban environmental management. These themes suggest opportunities for greater integration of AI with environmental governance, resource management, community resilience, and sustainability assessment. The SDG mapping (Fig. 2(G)) shows strong research emphasis on SDG 13 (2,941 publications), SDG 11 (1,963), and SDG 9 (1,346), reflecting priorities in climate action, urban resilience, and technological innovation. In contrast, lower representation of SDGs 1 (84), 5 (40), 10 (234), and 16 (188) indicates gaps in social inclusion, equality, governance, and institutional development. These gaps highlight the need for future research to emphasize governance, policy integration, community resilience, and socially relevant AI solutions.

4. Conclusions and Limitations

This study provides a bibliometric assessment of AI research on environmental sustainability, disaster risk reduction, and climate change resilience, identifying five major thematic clusters encompassing life cycle assessment, intelligent decision support, ecosystem monitoring, climate-risk prediction, and urban environmental sustainability. The strong representation of SDGs 13, 11, 9, 8, and 6 indicates substantial research attention to climate action, sustainable cities, technological innovation, and related sustainability challenges, while the limited emphasis on community resilience, policy integration, and governance highlights important research gaps. However, the study is subject to several limitations, including dependence on the Scopus database, which may exclude relevant publications indexed elsewhere; restriction to English-language articles, which may introduce language bias; and the use of Scopus indexing conditions, which may influence corpus composition and consequently the observed country, institutional, demographic, and SDG distributions. The 2026 records also represent partial-year coverage and should not be interpreted as a complete annual publication count. Furthermore, bibliometric analysis captures publication patterns, conceptual structures, and research priorities but cannot independently establish the operational effectiveness, reliability, environmental benefits, or societal impacts of AI applications. Future research should therefore complement bibliometric evidence with empirical investigation of AI use, governance, interdisciplinary collaboration, and social implications for sustainable development.

Acknowledgements

The authors acknowledge Amrita Vishwa Vidyapeetham for the support provided during the course of this work

References

- [1] T. T. Nguyen, U. Grote, F. Neubacher, M. H. Do, G. P. Paudel, et al., "Security risks from climate change and environmental degradation: implications for sustainable land use transformation in the Global South," *Current Opinion in Environmental Sustainability*, vol. 63, p. 101322, 2023.
- [2] K. Henderson and M. Loreau, "A model of Sustainable Development Goals: Challenges and opportunities in promoting human well-being and environmental sustainability," *Ecological Modelling*, vol. 475, p. 110164, 2023.
- [3] Z. Fan, Z. Yan, and S. Wen, "Deep learning and artificial intelligence in sustainability: a review of SDGs, renewable energy, and environmental health," *Sustainability*, vol. 15, no. 18, p. 13493, 2023.
- [4] S. U. Parvathy, V. K. Kolil, R. Raman, R. Vinuesa, and K. Achuthan, "Integrating sustainable development goals into life cycle thinking: a multidimensional approach for advancing sustainability," *Environment, Development and Sustainability*, pp. 1–39, 2025.
- [5] K. Achuthan, V. K. Kolil, S. U. Parvathy, N. R. Nambiar, and V. P. Shaw, "Life cycle assessment of coir geotextiles for sustainable infrastructure applications," *Journal of Cleaner Production*, vol. 554, p. 147879, 2026.
- [6] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, et al., "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *Revista Panamericana de Salud Pública*, vol. 46, p. e112, 2022.
- [7] A. Klarin, "How to conduct a bibliometric content analysis: Guidelines and contributions of content co-occurrence or co-word literature reviews," *International Journal of Consumer Studies*, vol. 48, no. 2, p. e13031, 2024.
- [8] G. E. Afandi, A. Moustafa, S. Ibrahim, and M. Irfan, "An in-depth analysis of approaches for forecasting extreme rainfall events leveraging artificial intelligence, machine learning, and deep learning techniques," *Earth Systems and Environment*, pp. 1–32, 2025.
- [9] A. C. Ikegwu, G. I. Emereonye, D. U. Ebem, and V. C. Uzuegbu, "A review of artificial intelligence expert systems for environmental surveillance and disaster management," *Discover Internet of Things*, vol. 6, no. 1, p. 26, 2026.
- [10] E. Alotaibi and N. Nassif, "Artificial intelligence in environmental monitoring: in-depth analysis," *Discover Artificial Intelligence*, vol. 4, no. 1, p. 84, 2024.
- [11] M. Al-Raei, "Artificial intelligence for climate resilience: advancing sustainable goals in SDGs 11 and 13 and its relationship to pandemics," *Discover Sustainability*, vol. 5, no. 1, p. 513, 2024.
- [12] S. S. Mishra, S. Sharma, Y. Sharma, and M. Singh, "Artificial intelligence for climate action and sustainable transitions: Conceptual frameworks, governance models, and responsible innovation pathways," in *AI for Climate Action: Governance, Justice, and Sustainable Transitions*, pp. 135–174, IGI Global Scientific Publishing, 2026.
- [13] F. Ullah, S. Saqib, and Y.-C. Xiong, "Integrating artificial intelligence in biodiversity conservation: Bridging classical and modern approaches," *Biodiversity and Conservation*, vol. 34, no. 1, pp. 45–65, 2025.
- [14] M. A. Mohammed, M. A. Ahmed, and V. H. Abdullayev, "Data-driven sustainability: Leveraging big data and machine learning to build a greener future," *Babylonian Journal of Artificial Intelligence*, vol. 2023, pp. 17–23, 2023.
- [15] N. Kazanskiy, R. Khabibullin, A. Nikonov, and S. Khonina, "A comprehensive review of remote sensing and artificial intelligence integration: Advances, applications, and challenges," *Sensors*, vol. 25, no. 19, Art. no. 5965, 2025.

The Emerging AI Power Crisis: Two Revolutions Collide

Imen Ghedhioui[†]

Operating Unit on Policy-Driven
Electronic Governance
United Nations University
Guimarães, Portugal
imen.ghedhioui@unu.edu

Morten Meyerhoff Nielsen[§]

Operating Unit on Policy-Driven
Electronic Governance
United Nations University
Guimarães, Portugal
meyerhoff@unu.edu

Soumaya Ben Dhaou[†]

Operating Unit on Policy-Driven
Electronic Governance
United Nations University
Guimarães, Portugal
bendhaou@unu.edu

João Martins[§]

Operating Unit on Policy-Driven
Electronic Governance
United Nations University
Guimarães, Portugal
joao.martins@unu.edu

Abstract

The convergence of artificial intelligence expansion and climate commitments presents a critical policy challenge: powering energy-intensive AI infrastructure. Across the Middle East and Africa, rapid AI data center expansion threatens climate commitments and exacerbates energy and water insecurity, challenging SDG 13 and SDG 12. Small Modular Reactors can offer dense, carbon-free baseload power for sovereign AI-enabled economies, yet its sustainable integration is undermined by unresolved land-use and water-management gaps, especially in arid regions. This policy analysis examines SMR-AI integration across four countries: Ghana, Saudi Arabia, South Africa, and the United Arab Emirates. Through comparative regulatory and technical analysis, we identify pathways for harmonizing nuclear safety licensing with data center certification, enabling cross-border energy sharing, and fostering inclusive access to sustainable high-performance computing infrastructure. The research contributes to the strategic vision of building AI infrastructure as instrument for achieving energy sovereignty, and climate commitments.

Keywords

Artificial Intelligence, Small Modular Reactors, Sustainable Development, Energy Governance, Digital Infrastructure

1 Introduction

The global Data Centre industry currently consumes an estimated 300-400 terawatt-hours of electricity annually, representing approximately 1-2% of worldwide power demand.[1]. This energy intensity is impacting electricity markets in key geographic hubs. Northern Virginia now hosts facilities consuming over 5 GW of power, a burden for the local grid capacity [3]. Still renewable procurement faces inherent

limitations: the intermittency of wind and solar cannot guarantee the 24/7 reliability that AI workloads demand. This is particularly significant in the context of the India-Middle East-Europe Economic Corridor (IMEC) and the emerging Digital Hub Corridor ambition [8].

2 Research Objectives and Methodology

2.1 Objectives

This research addresses the need for interdisciplinary analysis at the intersection of carbon-free nuclear energy governance, AI infrastructure, and sustainable development. The study aims to: (1) catalogue global strategies for integrating SMRs with AI data centers; (2) assess legal, regulatory, and technical barriers to deployment; and (3) co-create actionable policy recommendations for safe, sovereign nuclear-AI integration in two regional digital corridors through four selected case study countries: Saudi Arabia, UAE, Ghana, and South Africa.

2.2 Scope and Methodology

This research employs a mixed-methods approach combining: (1) regulatory and technical desk review of sources including national nuclear energy legislation, AI strategies, data protection laws, and cybersecurity frameworks; (2) comparative analysis of global nuclear-AI integration strategies; and (3) semi-structured stakeholder interviews with energy ministries, and international institutions.

3 The Global Energy-AI Nexus

3.1 Data Center Energy Demand and Infrastructure

According to the International Energy Agency, Data Centres consumed approximately 415 terawatt-hours of electricity in 2024, and projections indicate a potential to reach 945 TWh by 2030, determined primarily by AI [9].

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Information Society 2026, 5–9 October 2026, Ljubljana, Slovenia

© 2025 Copyright held by the owner/author(s).

<http://doi.org/10.70314/is.2026.ai4sdg.41>

3.2 Emerging Digital Corridors: The New Digital Geography of Risk

The next frontier of growth is the emerging global digital infrastructure markets with Africa growth for Data Centres and cloud computing, accounting for less than 1% of global Data Centre capacity despite 18% of the world's population, requires at least 1,000 MW of new capacity across 700 facilities [12]. Recent estimation, AI could contribute to the global economy, yet the Middle East is expected to receive only 2% [13].

3.3 The MENA Corridor: Geopolitical Energy Vulnerability, The Hormuz Factor and Regional Energy Security

The MENA Corridor faces a unique energy security challenge through its dependence on the Strait of Hormuz. Approximately 30% of the world's seaborne oil and 20% of global LNG pass through this chokepoint, making regional energy availability and pricing vulnerable to any disruption. [15].

3.4 The Africa Corridor: Structural Grid Vulnerabilities, The Load-Shedding Factor and Regional Energy Insecurity

4 The Small Modular Reactors For AI Infrastructure

4.1 SMR Technology Overview

Small Modular Reactors are advanced nuclear reactors with power capacity up to 300 MW(e) per unit approximately.

As of April 2025, two facilities meet commercial SMR criteria: Russia's Akademik Lomonosov and China's HTR-PM [25]. Over 80 SMR concepts are in design globally, with key developers including GE-Hitachi (BWRX-300), NuScale (VOYGR), Rolls-Royce, and Westinghouse. The BWRX-300, NuScale VOYGR, and SMART-100 have secured design approval from several national regulators [26].

5 Country's Analysis: Comparative Readiness Assessment

5.1 The United Arab Emirates

The UAE's Barakah plant comprising four APR-1400 reactors supplies 25% of national electricity, preventing 22.4 million tons of annual carbon emissions. ENEC signed strategic memoranda with Framatome and KEPCO to expand cooperation on SMRs and AI-driven digital transformation [29]. The 2025 Paris AI Action Summit catalyzed a framework accord for an AI-specific 1 GW Data Centre [30].

5.2 The Kingdom of Saudi Arabia

Saudi Arabia has emerged as one of the most strategically designed nuclear markets globally, with targeted institution-building before reactor selection embedded within Vision 2030. The separation of ownership through Saudi Nuclear Energy Holding Company from independent regulation via the Nuclear

and Radiological Regulatory Commission signals institutional readiness [31]. The U.S.-Saudi Arabia "Joint Declaration on Civil Nuclear Cooperation" in 2025 lays groundwork for a formal 123 Agreement [32].

5.3 South Africa

South Africa's Integrated Resource Plan 2025 mandates 5,200 MW of new nuclear capacity by 2039, with simultaneous revival of the Pebble Bed Modular Reactor program for indigenous SMR capabilities [33], [34]. South Africa has two operable reactors (4% nuclear share) but faces significant cybersecurity vulnerabilities (51.7 NCI), high carbon intensity (0.46), and grid collapse risks. The absence of data center-specific connection policies creates a governance vacuum. Without proactive grid policies that mandate on-site generation and demand flexibility for hyperscale facilities, South Africa risks compounding its existing load-shedding crisis with data center-driven grid strain. [35].

5.4 Ghana

Ghana secured plans to establish Africa's first NuScale Energy Exploration Center at the Ghana Atomic Energy Commission [36]. State-owned Nuclear Power Ghana targets operations of both an SMR and large reactor around 2034, with financing explored through build-own-operate-transfer models backed by government power purchase agreements [37]. Ghana has the highest renewable share (39.4%) among the four countries but no nuclear program and low AI readiness (0.43). Ghana's greenfield status offers a unique opportunity to embed on-site generation mandates, demand flexibility requirements, and grid impact assessment from the outset. [40].

6 Key Challenges and Barriers

The risk of a widening "sustainability divide" threatens to exclude countries from the global digital race [43]. As observed among the four use cases: Fragmented governance among energy, digital, nuclear, and cybersecurity authorities lacks joint oversight of the land-water-energy nexus, generating legal uncertainty around co-location siting and liability [41].

7 Policy Recommendations

7.1 Integrated Governance Frameworks

Governments should establish integrated energy-AI governance authorities with clear mandates. Regulatory frameworks must be adapted for modular, factory-fabricated designs, developing fit-for-purpose SMR licensing frameworks accounting for Data Centre colocation [44].

7.2 Technical Standards and interoperability

Technical standards must prioritize safety and interoperability, including cybersecurity architectures for SMR-data center integration, interoperability standards for energy sharing, and cooling technologies appropriate for local climate conditions. IEEE standards can enable integration between energy management systems and AI platforms [45].

7.3 Financing Model and Investment

Leveraging sovereign wealth funds for strategic, equity-based investments like the KSA and UAE cases can prioritize national ownership. Innovative mechanisms include build-own-operate-transfer models, Regulated Asset Base models, Corporate Power Purchase Agreements, and public-private partnerships balancing foreign expertise with local capacity [46].

7.4 Regional Cooperation

Regional bodies such as the African Union and Gulf Cooperation Council can support cross-border energy sharing, mutual safety certification, and shared technical standards. Regional interconnection grids can enable synergistic digital growth across borders [47].

7.5 Technology Negotiation and Capacity Building

Use cases showed the urge to develop strategies for effective negotiation with global technology providers, maintaining balanced cooperation over national AI strategy and critical data [48]. Building local capacity in nuclear engineering and AI development through education and partnerships is essential [49].

8 Conclusion

The coming decade presents a critical window for nations, particularly in the Global South, to shape their digital-energy sovereignty. With strategic governance, regional cooperation, and smart public-private partnerships, countries can harness SMRs to unlock their AI potential, establishing themselves as global models for a sustainable digital transition.

Acknowledgment

The authors thank UNU-EGOV and stakeholders who participated in interviews for their valuable contributions..

References

- [1] International Energy Agency, "Energy demand from AI," in Energy and AI, Paris, France: IEA, 2025. [Online]. Available: <https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai>
- [2] Global Electricity, "Data Centers and AI Energy Consumption: The Surge in Electricity Demand," Feb. 14, 2026. [Online]. Available: <https://www.globalelectricity.org/data-centers-energy-consumption/>
- [3] R. Mural, D. Pherwani, and C. Gupta, "Data Centers and Large-Scale Electric Growth: The Virginia and Texas Experiences," Belfer Center for Science and International Affairs, Harvard Kennedy School, Apr. 20, 2026. [Online]. Available: <https://www.belfercenter.org/research-analysis/data-centers-texas-virginia-comparison>
- [4] M. Peplow, "How the US tech industry is shaping the transition to green energy," *Nature*, vol. 639, no. 8055, pp. S18–S21, Mar. 2025. [Online]. Available: <https://www.nature.com/articles/d41586-025-00743-7>
- [5] "Constellation Energy to restart Three Mile Island nuclear plant, sell the power to Microsoft for AI," *CNBC*, Sep. 20, 2024. [Online]. Available: <https://www.cnbc.com/2024/09/20/constellation-energy-to-restart-three-mile-island-and-sell-the-power-to-microsoft.html>
- [6] M. Budaragina, D. West, A. Finotto Villena, and A. Phillips, "Powering AI in the Global South," Tony Blair Institute for Global Change, 2024. [Online]. Available: <https://institute.global/insights/climate-and-energy/powering-ai-in-the-global-south>
- [7] A. Kuehn and C. Ganapathy, "Building AI into the Energy Future," ORF Middle East, Aug. 5, 2026. [Online]. Available: <https://orfme.org/expert-speak/building-ai-into-the-energy-future/>
- [8] International Energy Agency, "Executive summary," in Energy and AI, Paris, France: IEA, 2025. [Online]. Available: <https://www.iea.org/reports/energy-and-ai/executive-summary>
- [9] A. Kuehn and C. Ganapathy, "Building AI into the energy future: Paradoxes and trendlines for the Global South," *Energy Reports*, vol. 15, pp. 115–130, Jun. 2026. doi: 10.1016/j.egyr.2026.05.001. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S221313882500342X>
- [10] D. Donnellan, A. Lawrence, D. Bizo, O. Rogers, P. Judge, J. Davis, M. Smolaks, and R. Weinschenk, "Uptime Institute Global Data Center Survey Results 2025 (Full Report)," Uptime Institute, 2025. [Online]. Available: <https://uptimeinstitute.com/resources/research-and-reports/uptime-institute-global-data-center-survey-results-2025>
- [11] World Economic Forum, "How shared digital infrastructure can bridge the gap in Africa," Apr. 9, 2025, Africa's data center capacity. [Online]. Available: <https://www.weforum.org/stories/economic-growth/how-shared-digital-infrastructure-can-bridge-the-gap-in-africa/>
- [12] L. Alajlouni and J. Murad, "The uncertain dividends of AI in the Middle East," *International Institute for Strategic Studies*, May 27, 2025. [Online]. Available: <https://www.iiss.org/online-analysis/charting-middle-east/2025/05/the-uncertain-dividends-of-ai-in-the-middle-east/>
- [13] A. Hussain and M. H. Kafafy, "Beyond the corridor: IMEC as a network of routes," Atlantic Council, Washington, DC, Jun. 2026. [Online]. Available: <https://www.atlanticcouncil.org/wp-content/uploads/2026/07/beyond-the-corridor-imec-as-a-network-of-routes.pdf>
- [14] U.S. Energy Information Administration, "Strait of Hormuz," in World Oil Transit Chokepoints, Washington, DC: U.S. EIA, Mar. 3, 2026. [Online]. Available: https://www.eia.gov/international/content/analysis/special_topics/World_Oil_Transit_Chokepoints/#1
- [15] V. Egoro, "Africa's \$120 Billion Grid Gap Is the Power Problem No One Is Fixing," *Energy Transition Africa*, Mar. 3, 2026. [Online]. Available: <https://www.energytransitionafrica.com/insights/article/africa-transmission-gap-grid-crisis-2026>
- [16] International Energy Agency, "Middle East," in World Energy Investment 2025, Paris: IEA, 2025. [Online]. Available: <https://www.iea.org/reports/world-energy-investment-2025/middle-east>
- [17] South Africa as an Investment Destination amid the Energy Crisis," *Journal of BRICS Studies*, vol. 4, no. 2, pp. 123–134, Jan. 2026. doi: 10.36615/hd54db53. [Online]. Available: <https://journals.uj.ac.za/index.php/jbs/article/view/2498>
- [18] K. Yeganegi, "The green transition in MENA economies: challenges and policy pathways," *Economic Research Forum*, Jun. 2, 2026. [Online]. Available: <https://theforum.erf.org/2026/06/02/the-green-transition-in-mena-economies-challenges-and-policy-pathways/>
- [19] D. Bahar and G. Wright, "New evidence on data center employment effects," *Brookings Institution*, Aug. 10, 2026. [Online]. Available: <https://www.brookings.edu/articles/new-evidence-on-data-center-employment-effects/>
- [20] J. W. Kane, "AI, data centers, and water: A growing need for regional coordination amid economic development potential," *Brookings Institution*, Nov. 20, 2025. [Online]. Available: <https://www.brookings.edu/articles/ai-data-centers-and-water/>

- [22] X. Chen, X. Wang, A. Colacelli, M. Lee, and L. Xie, "Electricity Demand and Grid Impacts of AI Data Centers: Challenges and Prospects," arXiv:2509.07218, Aug. 2026. [Online]. Available: <https://arxiv.org/html/2509.07218>
- [23] International Atomic Energy Agency, "What are Small Modular Reactors (SMRs)?," IAEA, Sep. 13, 2023. [Online]. Available: <https://www.iaea.org/newscenter/news/what-are-small-modular-reactors-smrs>
- [24] World Nuclear Association, "Advanced Nuclear Power Reactors," World Nuclear Association, Nov. 21, 2025. [Online]. Available: <https://world-nuclear.org/information-library/nuclear-power-reactors/other/advanced-nuclear-power-reactors>
- [25] "Renewable energy integration and energy efficiency in the Middle East and North Africa: A review of challenges and opportunities," Energy Reports, vol. 13, 2025. doi: 10.1016/j.egyr.2025.04.006. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2211467X2502044>
- [26] U.S. Department of Energy, "Advanced Small Modular Reactors (SMRs)," Office of Nuclear Energy, Jan. 30, 2013, Advanced SMR R&D Program. [Online]. Available: <https://www.energy.gov/ne/advanced-small-modular-reactors-smrs>
- [27] International Atomic Energy Agency, "Technology Roadmap for Small Modular Reactor Deployment," IAEA, Document Preparation Proposal Form VI Jan2020. [Online]. Available: <https://www.iaea.org/sites/default/files/20/08/technology-roadmap-for-small-modular-reactor-deployment.pdf>
- [28] EUREC - The Association of European Renewable Energy Research Centers, "Launch of the Strategic Roadmap on Digitalisation and AI in the Energy Sector and publication of ENTEC 2 study on Strategic Roadmap," Jun. 8, 2026. [Online]. Available: <https://eurec.be/launch-of-the-strategic-roadmap-on-digitalisation-and-ai-in-the-energy-sector-and-publication-of-entec-2-study-on-strategic-roadmap/>
- [29] Emirates Nuclear Energy Company, "ENEC and newcleo sign memorandum of strategic cooperation to explore developing nuclear projects globally," ENEC, Feb. 24, 2025. [Online]. Available: <https://www.enec.ae/news/latest-news/enec-and-newcleo-sign-memorandum-of-strategic-cooperation-to-explore-developing-nuclear-projects-globally/>
- [30] "Inside UAE's \$30 Billion Data Centre Investment to Fuel the AI Gold Rush," Data Center World Middle East, 2026. [Online]. Available: <https://www.datacenterworldmiddleeast.com/the-strategic-case-for-uae-data-centre-investment/>
- [31] World Nuclear Association, "Saudi Arabia - World Nuclear Outlook Report," World Nuclear Association, Jan. 19, 2026. [Online]. Available: <https://world-nuclear.org/our-association/publications/world-nuclear-outlook-report/saudi-arabia---world-nuclear-outlook-report>
- [32] A. Bollfrass, "The US–Saudi 123 Agreement: More than nuclear cooperation at stake," International Institute for Strategic Studies, Aug. 2026. [Online]. Available: <https://www.iiss.org/online-analysis/online-analysis/2026/08/the-ussaudi-123-agreement-more-than-nuclear-cooperation-at-stake/>
- [33] "South African government approves draft 2025 IRP," World Nuclear News, Oct. 22, 2025. [Online]. Available: <https://www.world-nuclear-news.org/articles/south-african-government-approves-draft-2025-irp>
- [34] South African Nuclear Energy Corporation, "South Africa will reactivate the PBMR small modular reactor project," Necs, Jul. 23, 2025. [Online]. Available: <https://www.necs.co.za/2025/07/23/south-africa-will-reactivate-the-pbmr-small-modular-reactor-project/>
- [35] "South Africa's Eskom targets data centre 'gold rush' for its power surplus," Financial Times, Aug. 12, 2026. [Online]. Available: <https://www.ft.com/content/107b354b-ba03-44f2-9939-517cbee11cd0>
- [36] "Ghana's NuScale framework agreement positions the country as a greenfield adopter with renewable leadership potential," in "Ghana's NuScale E2 Center," Africa Nuclear Business Platform, 2024. [Online]. Available: <https://africabusinessplatform.com/events/africa-nuclear-business-platform>
- [37] Nuclear Regulatory Authority of Ghana, "Nuclear Power," NRA Ghana, 2026. [Online]. Available: <https://nra.gov.gh/nuclear-power>
- [38] Ghana Atomic Energy Commission, "GAEC and Nuclear Power Ghana sign landmark MoU to drive nuclear energy development," GAEC, May 2024. [Online]. Available: <https://gaec.gov.gh/2024/05/>
- [39] World Nuclear Association, "Ghana - World Nuclear Outlook Report," World Nuclear Association, Jan. 19, 2026. [Online]. Available: <https://world-nuclear.org/our-association/publications/world-nuclear-outlook-report/ghana---world-nuclear-outlook-report>
- [40] E. Agyemang, "Ghana and Nuclear Energy: The Gap Between Promise and Readiness," IPPG Africa, May 4, 2026. [Online]. Available: <https://www.ippgafrica.org/blog-post/ghana-and-nuclear-energy-the-gap-between-promise-and-readiness/>
- [41] T. A. Downing and D. O. Olago, "History of 'nature-based' solutions to land degradation in the savannahs of Kenya: An analysis of literature and archival records," Environmental Science & Policy, vol. 161, p. 103874, Nov. 2024. doi: 10.1016/j.envsci.2024.103874.
- [42] M. G. Zamani, A. Moridi, and J. Yazdi, "Groundwater management in arid and semi-arid regions," Arabian Journal of Geosciences, vol. 15, no. 4, p. 362, Feb. 2022. doi: 10.1007/s12517-022-09546-w.
- [43] A. Lal and F. You, "Advances and challenges in energy and climate alignment of AI infrastructure expansion," Advances in Applied Energy, vol. 20, Art. no. 100243, Dec. 2025. doi: 10.1016/j.adapen.2025.100243.
- [44] Y. Cao, X. Li, H. Zhang, M. Zhai, H. Wu, C. Su, and R. Qi, "How the Transformation of Digital–Carbon Integration Is Empowering Sustainable Development: Theoretical Logic and Practical Pathways," Sustainability, vol. 18, no. 6, Art. no. 3159, Mar. 2026. doi: 10.3390/su18063159
- [45] M. Z. Khan, E. B. M. Bashir, M. M. Rana, and S. A. Hosseini, "Regulating AI in the Energy Sector: A Scoping Review of EU Laws, Challenges, and Global Perspectives," Energies, vol. 18, no. 9, Art. no. 2359, May 2025. doi: 10.3390/en18092359.
- [46] A. Al-Hassan and M. Al-Khalidi, "Sovereign Wealth Funds as Engines of Structural Economic Transformation and Income Diversification: A Comparative Mixed-Methods Analysis of Saudi Arabia's Public Investment Fund and Four Global Benchmark Funds," Preprints, vol. 202608.0431, Aug. 2026. doi: 10.20944/preprints202608.0431.v1.
- [47] World Bank, "Pushing the Frontiers of Clean Energy: Leveraging AI for a More Sustainable Future," World Bank, Washington, DC, Oct. 2025. [Online]. Available: <https://openknowledge.worldbank.org/server/api/core/bitstreams/bc5ac60-c902-485e-ad15-4e8bad9a201c/content>
- [48] International Telecommunication Union, "AI Standards for Global Impact: From Governance to Action," ITU, Geneva, Switzerland, Oct. 2025. [Online]. Available: <http://handle.itu.int/11.1002/pub/827794e7-en>
- [49] IAEA Collaborating Centre on Artificial Intelligence for Nuclear Power, "AI to Accelerate the Technological Development of Evolutionary and Innovative Reactor Designs and Fusion Facilities," Center for Science of Information at Purdue University, 2023. [Online]. Available: <https://unaihub.aiforgood.itu.int/activity-details.html?id=1601>

Decolonising the Digital Classroom: A Theoretical Case for AI-Mediated Indigenous Knowledge Integration in Open Distance eLearning

Dr. Terungwa James Age
Department of Mathematics Education
University of South Africa
Pretoria, South Africa
agetj@unisa.ac.za | ORCID: 0000-0001-7444-2213

Abstract

This conceptual paper addresses a specific problem in generative artificial intelligence (GenAI)-mediated open distance e-learning (ODEL): Indigenous Knowledge Systems (IKS) may become digitally visible while authority over their representation, interpretation and reuse remains external to Indigenous knowledge holders. It develops Epistemically Pluralist GenAI (EP-GenAI), integrating decolonial pedagogy, the capability approach, communicative rationality and Ubuntu into three conditions: ontological openness, epistemic co-authorship and curricular reciprocity. Indigenous Data Sovereignty supplies the governance condition of epistemic self-determination. The framework shifts evaluation from the amount of IKS included in GenAI to whether community-defined authority survives digital mediation.

Keywords

Generative AI, Indigenous Knowledge Systems, ODeL, Indigenous Data Sovereignty, epistemic plurality.

1 Introduction

The problem is not whether IKS can be integrated into GenAI-supported Open Distance eLearning (ODEL), but whether Indigenous communities retain authority once their knowledge is digitally represented. AI can support Indigenous knowledge initiatives while also enabling cultural erosion, extraction and disregard for Indigenous Data Sovereignty [1]. ODeL adds an institutional layer because platforms and curricula mediate what geographically dispersed learners encounter [2], while AI-driven educational infrastructures can reproduce epistemic exclusions [3].

Representation alone does not resolve this problem. The CARE Principles distinguish data availability from legitimate authority through Collective Benefit, Authority to Control, Responsibility and Ethics [4]. Empirical work on tribal knowledge also shows that GenAI can omit culturally important events and misrepresent tribal knowledge, while community participation improves the basis for evaluation [5]. The unresolved issue is therefore how representation, Indigenous authority and curricular interaction should be organised when GenAI mediates IKS.

This paper proposes Epistemically Pluralist GenAI (EP-GenAI). Its three conditions are ontological openness, epistemic

co-authorship and curricular reciprocity. These are theoretical propositions rather than properties attributed to current systems. Epistemic self-determination provides the governance condition connecting the three.

2 From inclusion to epistemic authority

Formal inclusion of IKS does not necessarily alter curricular hierarchy. Comparative work in South African and Australian higher education shows that moving beyond tokenism requires sustained engagement with Indigenous knowledge forms, including narrative and multimodal forms [6]. Decolonising curriculum research similarly cautions that adding Indigenous perspectives can leave dominant representational practices intact [7]. In ODeL, digitisation therefore raises two linked questions: whether knowledge form survives mediation and who determines legitimate representation.

GenAI adds linguistic asymmetry. Studies of multilingual and low-resource settings show representational bias and substantially weaker performance for languages with limited digital resources [8], [9]. For IKS, the response cannot simply be corpus expansion: more collection may conflict with community restrictions on what may be digitised, disclosed or reused. Indigenous Data Sovereignty therefore turns a technical data-gap problem into a governance problem [4], [10].

3 Theoretical basis

EP-GenAI assigns a specific function to each theoretical resource. Decolonial pedagogy identifies epistemic hierarchy and the persistence of dominant knowledge standards [3], [7]. Freirean dialogue rejects treating learners and knowledge holders as passive recipients or deposits of knowledge [11]. The capability approach adds substantive agency: participation matters when actors have real opportunities to influence valued outcomes, including curriculum and knowledge practices [12]–[14].

Communicative rationality contributes reciprocity by orienting interaction toward mutual understanding rather than predetermined authority [15]. Ubuntu adds relational responsibility: educational aims concern both self-realisation and the quality of relations through which collective life is sustained [16]. EP-GenAI uses these functions to specify what must change in representation, authority and curricular interaction rather than treating theoretical plurality as an end in itself.

4 EP-GenAI conceptual architecture

EP-GenAI translates the preceding problem into three evaluative conditions. Figure 1 shows their relationship. Epistemic self-determination underpins rather than supplements the three dimensions: without enforceable community authority, openness, co-authorship and reciprocity can remain symbolic.

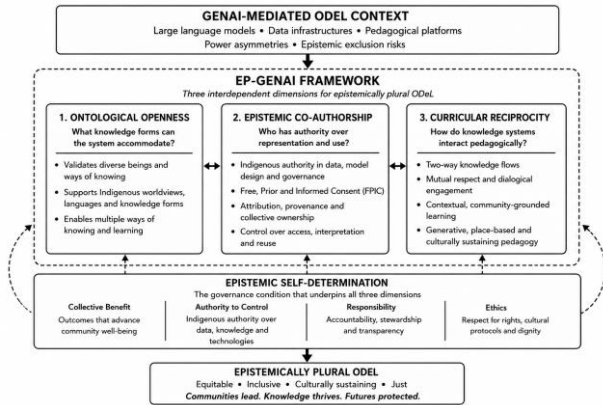


Fig. 1. EP-GenAI conceptual architecture. Epistemic self-determination governs the three framework dimensions.

4.1 Ontological Openness

Ontological openness asks whether digital mediation can preserve different knowledge forms without forcing them into one representational logic. It does not mean unrestricted digitisation. Appropriate implementation must preserve provenance, context and restrictions, including non-representation where knowledge is not authorised for disclosure. The criterion is whether the system preserves conditions that make knowledge intelligible and legitimate, not merely whether an LLM can generate text about IKS.

4.2 Epistemic Co-authorship

Epistemic co-authorship concerns decision authority. Knowledge holders should influence what enters an OER or GenAI workflow, how material is represented and attributed, and what access or reuse conditions apply. This exceeds consultation: contribution is insufficient if an institution or provider independently controls subsequent processing. Tribal knowledge co-creation provides empirical support for involving communities in the sources and processes used to improve GenAI representation [5].

4.3 Curricular Reciprocity

Curricular reciprocity concerns relations among knowledge systems. It does not require treating different claims as interchangeable. It requires learning activities to preserve differences in evidence, context and purpose rather than silently positioning one tradition as the universal interpretive standard. IKS can then shape how educational problems and technologies are framed, rather than functioning only as content processed by AI [1].

5 Epistemic self-determination

Indigenous Data Sovereignty supplies EP-GenAI's governance condition. CARE places Indigenous rights and interests within data stewardship and reuse [4], while analyses of data mining show how extraction can reproduce colonial relations when Indigenous peoples lack control over data practices [10]. Epistemic self-determination therefore means substantive community authority over representation, access, attribution, interpretation and reuse.

This distinction matters for open education. An OER may be accessible for learning without authorising model training, unrestricted generation or secondary datasets. Community-defined restrictions must survive movement from a learning resource into GenAI infrastructure. Where a platform cannot preserve access, retention or reuse conditions, the institution cannot claim that community authority has been maintained.

6 Implications for ODeL

EP-GenAI yields four operational requirements. First, ODeL resources should preserve appropriate textual, oral or multimodal forms rather than converting all IKS into decontextualised text. Second, provenance and community-defined access conditions should remain attached to resources. Third, GenAI outputs should identify the epistemic location of accounts when synthesising knowledge traditions. Fourth, procurement and deployment should be conditional on whether institutions can enforce restrictions on retention, reuse and model development.

These requirements make governability, rather than content quantity, the adoption criterion. A course containing extensive IKS can still fail EP-GenAI if knowledge holders cannot correct representation, restrict reuse or withdraw material. Conversely, limited digital representation may be appropriate where community protocols require it.

7 Discussion

EP-GenAI does not claim that better prompts or larger Indigenous datasets can decolonise current GenAI. Indigenous-AI scholarship documents extraction and algorithmic settler-colonial dynamics that extend beyond interface design [1], [17]. Foundation-model research also locates harms in labour and natural-resource extraction, not only generated content [18]. Epistemic justice, therefore, cannot be assessed solely from the inclusiveness of outputs.

The framework instead establishes four tests: whether representation preserves epistemic form and legitimate restrictions; whether knowledge holders possess decision authority; whether curricular encounters avoid predetermined hierarchy; and whether these arrangements remain governable by the relevant communities. Failure on the fourth test limits claims of epistemic democratisation even when generated outputs appear culturally inclusive.

The distinction between representation and authority is especially important for model outputs. Even where Indigenous material is present in a training or retrieval corpus, generation

can detach claims from provenance, compress differences among traditions, or present a synthesis as if it were epistemically neutral. EP-GenAI therefore treats output interpretation as part of ontological openness. A system should not be judged open merely because it accepts diverse inputs; it must also avoid transforming those inputs in ways that erase their source, status or restrictions.

Co-authorship has a corresponding infrastructural threshold. A community may help design an ODeL resource yet remain unable to determine how a commercial model stores prompts, reuses uploaded material or incorporates interactions into later development. In that case, participation exists at the content layer but not at the infrastructure layer. Roberts and Montoya's account of data mining as a colonial practice [10] clarifies why this distinction matters: control over collection without control over subsequent processing does not amount to data sovereignty.

Curricular reciprocity also requires more than presenting two accounts side by side. GenAI systems routinely summarise and reconcile information, but reconciliation can conceal genuine epistemic difference. In an EP-GenAI learning activity, the model should preserve differences in the grounds on which claims are warranted, identify whose account is being presented, and avoid converting disagreement into a single authoritative synthesis. Reciprocity therefore concerns the structure of comparison, not a requirement that all knowledge claims receive identical validation.

These conditions change how ODeL institutions should evaluate AI services. Model accuracy remains relevant, but it is insufficient for Indigenous knowledge integration. Evaluation should also ask whether provenance can be retained, whether restricted material can be excluded from secondary use, whether communities can amend or withdraw representations, and whether contractual terms permit institutional enforcement of those decisions. The framework consequently links pedagogical design to procurement and data governance rather than treating them as separate administrative matters.

The framework also places a boundary on claims of decolonisation. Worrell and Carlson [17] show that Indigenous critiques of AI extend to algorithmic settler-colonial relations, while foundation-model research documents labour and natural-resource extraction across AI supply chains [18]. An interface may therefore appear culturally responsive while depending on infrastructures whose governance conflicts with the relational and sovereignty commitments invoked by EP-GenAI. The framework does not resolve that contradiction; it makes it visible as an evaluative constraint.

For ODeL practice, this means that non-use can sometimes be a valid outcome. If a community does not authorise digitisation of particular knowledge, or if a platform cannot honour restrictions on retention and reuse, epistemic openness does not require inclusion. Respecting a boundary can be more consistent with epistemic self-determination than maximising the amount of Indigenous content available to learners. This point prevents EP-GenAI from equating plurality with unrestricted access.

8 Limitations and future research

EP-GenAI is conceptual and does not demonstrate that current ODeL platforms can satisfy its three dimensions simultaneously. It also cannot specify which knowledge may be digitised or who may authorise its use across diverse Indigenous communities. A further limitation is infrastructural: community restrictions may not be technically enforceable in externally controlled GenAI platforms [17].

Future research should test community-governed ODeL cases by examining whether restrictions survive GenAI processing, whether knowledge holders can correct or withdraw material, whether provenance remains visible in outputs, and whether different knowledge traditions retain distinct epistemic positions in learning activities.

9 Conclusion

EP-GenAI reframes Indigenous knowledge integration in GenAI-mediated ODeL as a problem of representation, authority and curricular relation. Ontological openness, epistemic co-authorship and curricular reciprocity operate under epistemic self-determination. The framework is evaluative rather than technical: its empirical test is whether community-defined conditions can be implemented and enforced within real ODeL infrastructures.

References

- [1] M. Perera et al., "Indigenous peoples and artificial intelligence: A systematic review and future directions," *Big Data & Society*, 2025, <https://10.1177/20539517251349170>
- [2] V. J. Pitsoe, "Interrogating epistemic foundations: Open distance e-learning for inclusive knowledge production," *Zenodo*, 2025, <https://10.5281/zenodo.15250666>
- [3] M. Popow, "Decolonizing knowledge in the postdigital era: Pedagogical strategies for navigating AI-driven epistemic transformations," *Educational Philosophy and Theory*, 2025, <https://10.1080/00131857.2025.2552764>
- [4] S. R. Carroll et al., "The CARE Principles for Indigenous Data Governance," *Data Science Journal*, vol. 19, Art. no. 43, 2020, <https://10.5334/dsj-2020-043>
- [5] Y.-F. Wang et al., "Tribal knowledge cocreation in generative artificial intelligence systems," *Proc. CHI Conf. Human Factors Comput. Syst.*, 2024, <https://10.1145/3657054.3657129>
- [6] M. Keane et al., "Indigenous Knowledge Systems in South Africa and Australia: Transforming doctoral education," *Curriculum Perspectives*, 2023, <https://10.1007/s41297-023-00183-1>
- [7] N. Harrison and I. Clarke, "Decolonising curriculum practice: Developing the Indigenous cultural capability of university graduates," *Higher Education*, 2021, <https://10.1007/s10734-020-00648-6>
- [8] D. I. Adelani et al., "Which Nigerian-Pidgin does Generative AI speak?: Issues about representativeness and bias for multilingual and low resource languages," *arXiv:2404.19442*, 2024, <https://10.48550/arXiv.2404.19442>
- [9] C. Chaka, "Currently available GenAI-powered large language models and low-resource languages: Any offerings? Wait until you see," *Int. J. Learning, Teaching Educ. Research*, vol. 23, no. 12, 2024, <https://10.26803/ijlter.23.12.9>

- [10] J. S. Roberts and L. N. Montoya, "In consideration of Indigenous Data Sovereignty: Data mining as a colonial practice," arXiv:2309.10215, 2023, <https://10.48550/arXiv.2309.10215>
- [11] P. Freire, *Pedagogy of the Oppressed*. New York, NY, USA: Continuum, 1970
- [12] D. Talbot, "Knowledge, knowers, and capabilities: Can the capabilities approach help decolonise the curriculum?" *J. Human Development and Capabilities*, 2023, <https://10.1080/19452829.2023.2200240>
- [13] M. Walker, "Amartya Sen's capability approach and education," *Educational Action Research*, vol. 13, 2005, doi: <https://10.1080/09650790500200279>
- [14] A. Sen, *Development as Freedom*. New York, NY, USA: Alfred A. Knopf, 1999.
- [15] J. Habermas, *The Theory of Communicative Action*, vol. 1, T. McCarthy, Trans. Boston, MA, USA: Beacon Press, 1984.
- [16] C. H. Manthalu, "A sketch of an Ubuntu philosophy of education," 2023, https://10.1007/978-3-031-25149-8_23
- [17] T. Worrell and B. Carlson, "Indigenous AI futures: Uncle Chatty Gee, Aunty Lexi, and algorithmic settler colonialism," *Somatechnics*, 2025, <https://10.3366/soma.2025.0468>
- [18] A. Domínguez Hernández et al., "Mapping the individual, social and biospheric impacts of Foundation Models," *Proc. ACM FAccT*, 2024, <https://10.1145/3630106.3658939>

AI-Enabled Knowledge Infrastructures for Social Innovation: Business Value, Adoption, and Governance

Syed Muhammad Ali Abbas
Syed.Muhammad.Ali.Abbas@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Matjaž Gams
matjaz.gams@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

Social innovation knowledge is distributed across academic publications, policy documents, project repositories, non-governmental organisations, and local initiatives. This fragmentation creates a practical business problem for public agencies, social enterprises, research organisations, funders, and consultancies: evidence discovery is slow, successful practices are difficult to compare, and knowledge is repeatedly recreated rather than reused. This paper reframes artificial-intelligence-enabled knowledge infrastructures as both a research direction and an organisational capability. Drawing on work in social innovation, AI for Social Good, digital social innovation, retrieval-augmented generation, and knowledge graphs, it identifies four sources of value: lower evidence-search costs, faster decision support, improved transfer of proven practices, and stronger ecosystem coordination. A conceptual architecture is proposed that combines structured extraction, hybrid retrieval, source-grounded generation, and human validation. The paper further outlines potential operating models, cost drivers, adoption barriers, governance requirements, and measurable business outcomes. The central argument is that the value of AI in social innovation lies not only in model performance, but in converting fragmented knowledge into a trustworthy and reusable organisational asset.

Keywords

artificial intelligence, social innovation, business value, knowledge infrastructure, large language models, retrieval-augmented generation, knowledge graphs, responsible AI

1 Introduction

Social innovation addresses unmet societal needs through new practices, relationships, services, and institutional arrangements rather than through technology alone [4, 10]. Relevant knowledge is consequently dispersed across research disciplines, policy reports, project databases, NGO documents, community initiatives, and multilingual sources. Existing repositories provide useful documentation, but they often remain static and difficult to compare. For organisations that fund, design, evaluate, or scale social innovation, this fragmentation produces real operational costs: analysts spend substantial time locating evidence, similar projects are duplicated, lessons are lost between programmes, and decision makers struggle to judge whether an intervention is transferable to a new context.

Recent progress in large language models (LLMs), retrieval-augmented generation (RAG), semantic search, and knowledge

graphs (KGs) creates an opportunity to treat this problem as a knowledge-infrastructure challenge. Instead of deploying a generic chatbot over a document collection, organisations can construct systems that extract structured information, connect cases and stakeholders, retrieve evidence, preserve provenance, and support human validation. This direction builds on digital social innovation [8], AI for Social Good [2], Data Science for Social Good (DSSG) [1], and research on the application of AI to social innovation [6].

Beyond its organisational value proposition, the proposed framework also has a sustainable-development rationale. Social-innovation organisations frequently operate across institutional boundaries and depend on fragmented evidence to design, justify, adapt, and coordinate interventions. A provenance-aware knowledge infrastructure can strengthen these processes by making evidence more traceable, comparable, and reusable, thereby supporting more transparent and accountable institutional decision-making. This aligns most directly with SDG 16, particularly Target 16.6 on effective, accountable, and transparent institutions, while the framework's emphasis on stakeholder participation and continuous validation also relates to Target 16.7 on inclusive and participatory decision-making [13]. The contribution is enabling rather than direct: the infrastructure does not itself produce sustainable-development outcomes, but can strengthen the institutional conditions under which public agencies, NGOs, funders, and social enterprises make and coordinate evidence-informed decisions.

H1: Compared with conventional repository-based approaches, AI-enabled knowledge infrastructures that combine structured knowledge representation, source-grounded retrieval, and human validation are expected to reduce evidence-search effort and improve the reuse of social innovation knowledge for organisational decision support. H1 is proposed for subsequent empirical validation (Section 6).

This paper integrates these research streams into a domain-specific conceptual framework for AI-enabled social-innovation knowledge infrastructures. It makes three contributions: (1) it combines heterogeneous social-innovation evidence, structured knowledge representation, provenance-aware retrieval and generation, and continuous human validation within a unified knowledge-infrastructure framework; (2) it identifies the organisational value mechanisms and operating models through which such infrastructures may support evidence discovery, knowledge reuse, decision support, and ecosystem coordination; and (3) it links technical design with adoption, governance, and measurable technical, organisational, and social outcomes. The contribution is conceptual: it provides a structured basis for designing and evaluating future systems rather than claiming an empirically validated solution.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Information Society 2026, Ljubljana, Slovenia

© 2026 Copyright held by the owner/author(s).

<https://doi.org/10.70314/is.2026.ai4sdg.38>

2 Review and Research Gap

This conceptual study uses a purposive, narrative synthesis rather than a systematic literature review. The literature was selected to represent five complementary streams relevant to the research problem: social innovation, digital social innovation, AI and Data Science for Social Good, AI-enabled knowledge technologies such as retrieval-augmented generation and knowledge graphs, and responsible-AI governance. These streams were compared according to the capabilities they provide and the organisational needs that remain unresolved (Table 1). The proposed framework was then derived by mapping these gaps to four connected design dimensions: knowledge representation and retrieval, organisational value creation, operating and adoption requirements, and governance and evaluation. This synthesis provides the conceptual basis for the architecture, value mechanisms, and evaluation approach developed in the following sections.

Social innovation research emphasises that value is created through changes in practices, relationships, and institutional arrangements rather than through isolated technological artefacts. Mapping initiatives such as the Atlas of Social Innovation have demonstrated the diversity of actors and contexts involved, but they also expose the difficulty of comparing cases across sectors and regions [4]. Digital social innovation extended this field through platforms, open data, civic technologies, and online collaboration. These systems improved access and participation, yet many retained a repository-oriented design in which information was stored and indexed but not semantically connected or actively synthesised [8].

AI for Social Good and Data Science for Social Good demonstrate how computational methods can be directed towards public-interest objectives [2, 1]. Their main contribution is application-oriented: prediction, classification, optimisation, and decision support in domains such as health, education, climate action, poverty reduction, and public services. AI for Social Innovation requires a broader organisational perspective. A technically successful model does not automatically create social or business value. Value depends on whether organisations can embed the system in workflows, interpret its outputs, govern its risks, retain knowledge, and improve decisions over time.

LLMs, retrieval-augmented generation, and knowledge graphs address important parts of this problem by supporting structured extraction, semantic retrieval, relationship modelling, and evidence-grounded generation. However, these capabilities usually address only individual layers of the knowledge problem. Repository-oriented systems primarily support storage and access; RAG systems improve retrieval and generation; knowledge graphs provide structured relationships; and responsible-AI frameworks emphasise accountability, transparency, and human oversight [12, 14, 9]. What remains underdeveloped in the social-innovation context is an integrated knowledge-infrastructure framework that connects these technical capabilities with heterogeneous evidence sources, provenance, structured domain knowledge, continuous human and stakeholder validation, organisational value creation, governance responsibilities, and measurable outcomes. The contribution of this work therefore lies not in proposing a new AI technique, but in integrating these elements into a domain-specific framework for the design, operation, and evaluation of AI-enabled social-innovation knowledge infrastructures (Table 1).

Table 1: Research streams and unresolved organisational needs

Stream	Main contribution	Remaining organisational need
Social innovation	Cases, ecosystems, practices	Comparable and reusable evidence
Digital social innovation	Platforms, participation, open data	Semantic links and active synthesis
AI/DSSG	Prediction and optimisation for public value	Workflow integration and impact measures
LLM/RAG/KG	Extraction, retrieval, generation	Provenance, governance, and sustainable operation
Responsible AI governance	Accountability, transparency, human oversight	Domain-specific governance and continuous stakeholder participation

3 From Knowledge Fragmentation to Business Value

The business problem is not simply a shortage of information. It is the cost of converting heterogeneous information into reliable decisions. Public agencies need evidence for policy design; funders need portfolio visibility and due diligence; NGOs and social enterprises need transferable practices; universities and consultancies need efficient synthesis; and innovation networks need shared vocabularies and ecosystem maps. Social enterprises in particular face opportunities and constraints when adopting AI, including limited resources, capability gaps, organisational readiness, and the need to preserve mission alignment [11].

An AI-enabled knowledge infrastructure may create value through four mechanisms. First, it can reduce search and synthesis effort by retrieving relevant evidence across scholarly and practice-oriented sources. Second, it can accelerate decision support through source-grounded comparisons of interventions, actors, methods, costs, barriers, and outcomes. Third, it can support organisational learning by preserving knowledge beyond individual projects and staff turnover. Fourth, it can improve ecosystem coordination by making relationships between funders, implementers, regions, target groups, and outcomes more visible.

Table 2 translates these mechanisms into stakeholder-specific outcomes. The proposed metrics deliberately extend beyond model accuracy. They include time to evidence, analyst hours saved, reuse of prior cases, decision lead time, and user adoption. These indicators are closer to the operational value that organisations expect from a knowledge product.

4 Architecture and Operating Model

The proposed architecture (Figure 1) follows a traceable data flow from heterogeneous evidence sources to organisational use. Academic publications, policy documents, project repositories, and practice-based sources are first ingested and enriched with metadata and provenance. LLM-assisted structured extraction then maps relevant content to domain fields such as social problem, intervention, stakeholder, method, evidence, limitation, outcome, and transfer condition. These representations are stored in a hybrid knowledge layer combining taxonomies, graph relationships, metadata, and vector embeddings. Retrieval combines semantic similarity with structured metadata and graph relationships to identify relevant evidence before generation [7, 5]. Generated

Table 2: Stakeholders, value propositions, and business measures

Stakeholder	AI-enabled value	Illustrative measure
Public agencies	Faster policy evidence and comparison of interventions	Decision lead time; evidence coverage
NGOs and social enterprises	Access to transferable practices, partners, and funding evidence	Search time; case reuse; adoption
Funders and networks	Portfolio mapping, due diligence, and duplication detection	Review effort; overlap identified
Research and consulting organisations	Automated extraction, synthesis, and gap analysis	Analyst hours saved; report cycle time

summaries and comparisons remain linked to their supporting sources so that claims can be inspected and validated. Human experts and community stakeholders participate not only in final review, but also in source selection, metadata and taxonomy design, correction of extracted knowledge, and periodic validation. Their feedback is incorporated back into the knowledge layer, creating a continuous improvement loop rather than a one-time validation step.

The infrastructure can be delivered through several operating models. A subscription-based software-as-a-service (SaaS) model may suit consultancies, research organisations, or large NGOs that require private workspaces and integrations. A consortium or publicly funded model may be more appropriate where knowledge is considered shared infrastructure and broad accessibility is a policy objective. A hybrid model can combine an open core with paid services for data ingestion, custom taxonomies, private deployments, analytics, and training. The appropriate model depends on data sensitivity, funding structure, user diversity, and the degree to which the infrastructure is treated as a public good.

Major cost and sustainability drivers include corpus licensing, document processing and curation, LLM and embedding inference and associated energy use, storage, expert validation, security, integration, and taxonomy maintenance. These costs should be assessed against avoided analyst effort, reduced duplication, improved programme design, and faster access to evidence. A staged implementation limits risk: an initial minimum viable product can focus on one social domain and a curated corpus; a pilot can then test retrieval quality and user value; later phases can expand to multilingual and practice-based sources.

5 Commercialisation and Staged Deployment

A sustainable deployment requires a clear answer to who owns the infrastructure, who pays for it, and which functions remain shared. Four operating configurations are plausible (Table 3). A commercial software-as-a-service product can provide private workspaces, access control, connectors, analytics, and service-level agreements. A publicly funded or consortium model can treat the corpus, taxonomy, and evidence layer as shared infrastructure. An open-core model can make basic retrieval publicly accessible while charging for private deployments, integrations, monitoring, and specialist support. Finally, a professional-services model can begin with project-based evidence synthesis before reusable platform components are standardised.

Table 3: Illustrative operating models

Model	Best suited to	Funding or revenue logic
SaaS	Large NGOs, consultancies, research organisations	Subscription by users, corpus size, or integrations
Public or consortium	Governments, universities, innovation networks	Grants, membership fees, or shared public funding
Open core	Mixed public and private ecosystem	Free core; paid hosting, private data, and support
Professional services	Early pilots and specialised domains	Project fees for curation, analysis, and implementation

The economic case should be tested through total cost of ownership rather than model price alone. Direct costs include data acquisition, document conversion, storage, embeddings, LLM inference, identity management, security, monitoring, and expert review. Indirect costs include taxonomy maintenance, user training, integration with existing repositories, and governance work. Benefits should be estimated against a baseline workflow: hours spent searching and screening documents, cost of duplicated analysis, delays in programme design, and the value of reusing evidence across projects. This makes the business case measurable even when the infrastructure is funded as a public good rather than operated for profit.

A staged deployment reduces both technical and market risk. Stage one should build a minimum viable corpus for one domain, such as social-care innovation or digital inclusion, and support traceable search and structured case comparison. Stage two should test the system with a small set of users and measure relevance, time saved, trust, and repeated use. Stage three can add private organisational data, multilingual sources, dashboards, and integrations. Only after these steps should the infrastructure expand across domains. This progression separates evidence of user value from the temptation to scale a broad but weakly validated platform.

From a product perspective, defensibility is more likely to come from the curated corpus, taxonomy, validated extraction pipeline, workflow integration, and trusted community than from access to a general-purpose LLM. Models can be replaced; accumulated domain knowledge, provenance links, and institutional adoption are harder to reproduce. The long-term asset is therefore the governed knowledge layer and the network of contributors who maintain it.

6 Adoption, Governance, and Evaluation

Technical feasibility does not guarantee organisational adoption. Users must trust generated claims, understand system limitations, and see clear benefits relative to existing workflows. Adoption barriers include poor source quality, limited AI skills, integration effort, privacy concerns, ownership uncertainty, vendor lock-in, and variable inference costs. Governance should therefore span the knowledge lifecycle. Provenance records should capture source, licensing, and access conditions; sensitive organisational or community data should use data minimisation and access controls; and corpus design should monitor representation gaps in local, minority-language, and under-documented knowledge. Domain experts and affected communities should participate in source selection, taxonomy design, validation, and periodic review, consistent with responsible-innovation guidance [12],

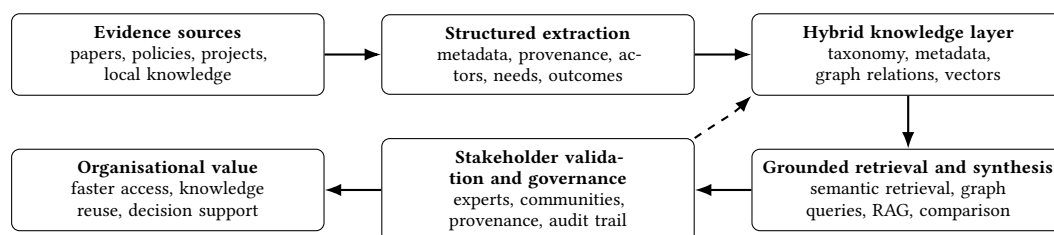


Figure 1: Proposed AI-enabled social-innovation knowledge infrastructure with stakeholder feedback to the knowledge layer.

human-oversight expectations [14, 9], and applicable AI Act obligations [3].

Evaluation should compare the proposed infrastructure with a baseline repository or search workflow on representative evidence-discovery and case-comparison tasks. Four layers are assessed: technical performance (retrieval relevance, faithfulness, accuracy, latency); knowledge quality (coverage, taxonomy usefulness, traceability); organisational value (search time, analyst effort, decision lead time, reuse, adoption); and social governance (participation, multilingual inclusion, bias, accountability, context sensitivity). A pilot with representative users would compare these outcomes under baseline and system-assisted workflows, providing an empirical test of H1.

7 Discussion and Conclusion

AI-enabled knowledge infrastructures offer a plausible business and organisational response to fragmentation in social innovation. Their value proposition is strongest where decisions depend on dispersed evidence, repeated comparison, and coordination across many actors. The infrastructure can reduce the transaction costs of finding, interpreting, and reusing knowledge, while creating a persistent organisational memory that survives individual projects.

However, the business case remains conceptual. Claims about efficiency, cost reduction, and decision quality require validation with real users and workflows. The most appropriate next step is a domain-specific pilot that measures baseline search effort, retrieval quality, user trust, and time-to-decision before and after deployment. A narrow, validated system is more credible than a broad platform that collects large volumes of poorly governed content.

The central conclusion is that AI creates business value in social innovation when it transforms fragmented evidence into a trustworthy, reusable, and governed knowledge asset. LLMs, RAG, and knowledge graphs are enabling components, not the final product. Sustainable adoption depends on aligning technical architecture with stakeholder needs, operating costs, governance responsibilities, and measurable organisational outcomes.

Acknowledgements

This paper was produced as part of the Datavysts MSCA network, funded by Horizon Europe [grant number 101169037] and UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding Guarantee [grant number EP/Z536581/1].

References

- [1] Ahmed Abbasi, Roger H. L. Chiang, and Jingjing Xu. 2023. Data science for social good. *Journal of the Association for Information Systems*, 24, 6, 1439–1458. doi:10.17705/1jais.00849.
- [2] Josh Cows, Andreas Tsamados, Mariarosaria Taddeo, and Luciano Floridi. 2021. A definition, benchmark and database of ai for social good initiatives. *Nature Machine Intelligence*, 3, 111–115. doi:10.1038/s42256-021-00296-0.
- [3] European Union. 2024. Artificial intelligence act. European Parliament and Council of the European Union. (2024).
- [4] Jürgen Howaldt, Christoph Kaletka, Antonius Schröder, and Marthe Zirngiebl, eds. 2018. *Atlas of Social Innovation: New Practices for a Better Future*. Sozialforschungsstelle, TU Dortmund University, Dortmund, Germany.
- [5] Mohamad Yaser Jaradeh, Allard Oelen, Manuel Prinz, Markus Stocker, and Sören Auer. 2022. Open research knowledge graph: a system walkthrough. *arXiv preprint arXiv:2206.01439*.
- [6] E. Kim, G.-Y. Jang, and S.-H. Kim. 2022. How to apply artificial intelligence for social innovations. *Applied Artificial Intelligence*, 36, 1. doi:10.1080/08839514.2022.2031819.
- [7] Patrick Lewis et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*. Vol. 33, 9459–9474.
- [8] J. Millard and G. Carpenter. 2014. Digital Technology in Social Innovation: A Synopsis. Tech. rep. TEPSIE Project, European Commission FP7.
- [9] OECD. 2024. Governing with Artificial Intelligence. Tech. rep. OECD Publishing, Paris, France.
- [10] A. Phillips, R. Luo, and J. Wendland-Liu. 2024. Shifting the paradigm: a critical review of social innovation literature. *International Journal of Innovation Studies*, 8, 1, 45–58. doi:10.1016/j.ijis.2023.08.003.
- [11] Artur Steiner, Micaela Mazzei, Francesca Calò, and Gordon Liu. 2025. How and why social enterprises adopt artificial intelligence: opportunities, challenges, and future implications for research and practice. *Journal of Social Entrepreneurship*. doi:10.1080/19420676.2025.2584843.
- [12] Jack Stilgoe, Richard Owen, and Phil Macnaghten. 2013. Developing a framework for responsible innovation. *Research Policy*, 42, 9, 1568–1580. doi:10.1016/j.respol.2013.05.008.
- [13] United Nations. 2015. Transforming our world: the 2030 agenda for sustainable development. (2015). <https://sdgs.un.org/2030agenda>.
- [14] World Economic Forum. 2024. AI for Impact: The PRISM Framework for Responsible AI in Social Innovation. Tech. rep. World Economic Forum and Schwab Foundation for Social Entrepreneurship, Geneva, Switzerland.

AIDE: Leadership Enabling AI Learning in Marginalised Communities through HCI

Mohammad Rizal Mohd Yusuf

Research Fellow
International Research Center
Artificial Intelligence (IRCAI)
Malaysia
rizal.yusuf@moe.gov.my

Julieyana Jaris

PhD Researcher
Universiti Teknikal Malaysia Melaka
(UTeM)
Malaysia
julieyana@moe.gov.my

Abstract / Povzetek

This study reports AI literacy achievement evidence from the Artificial Intelligence Decomplexify Education (AIDE) Model for marginalised communities and embeds AI Literacy Leadership (ALL) within AIDE as its institutional enabling layer. This modification follows the accepted leadership-SDG framing in which school or community leaders are not peripheral stakeholders but gatekeepers and catalysts of access: without positive leadership response, authorisation, mobilisation and sustained support, marginalised participants may not enter or remain in AI-learning programmes. The Middleware AI-Brain (MAI-BRAIN) Framework provides the conceptual foundation; Design and Development Research (DDR) produced AIDE as the model outcome; and Rural Innovation Making Better Agriculture (RIMBA) operationalised AIDE as a proof-of-concept with Indigenous Mah Meri and rural FELDA Gedangsa cohorts. Evidence combines task performance, post-intervention questionnaire data, descriptive statistics, reliability, content validity, qualitative feedback and supplementary pre-post analysis. Mah Meri achieved 18/20 (0.90) independent visual-programming completion, while FELDA Gedangsa achieved approximately 0.99 success against defined collaborative implementation criteria. Tangible-to-Digital Mapping (M=4.08), Multisensory Feedback (M=4.21) and Contextual IoT Bridging (M=4.34) were rated highly ($\alpha=.78-.83$; S-CVI/Ave=.92). Supplementary algorithmic-rule scores increased from M=1.55 to M=4.50, $t(39)=27.54$, $p<.001$, $d_z=4.35$. Leadership is therefore modelled as the institutional condition that authorises access, motivates participation, mobilises resources, supports facilitators, governs AI responsibly and catalyses translation of new knowledge into SDG-aligned social and economic capability.

Keywords / Ključne besede

AIDE; AI Literacy Leadership; HCI; MAI-BRAIN; marginalised communities; RIMBA; SDG

1. Conceptual and Development Positioning

The development chain remains MAI-BRAIN → DDR → AIDE → RIMBA. MAI-BRAIN integrates cognitive and multimedia processing, HCI and technology acceptance, motivation and engagement, and situated learning [2]-[5]. DDR operationalised these foundations into AIDE [1], while RIMBA provided authentic field implementation. The leadership-SDG framing makes an important implementation requirement explicit: leadership must be embedded within AIDE rather than appended after the learner-level model. This positioning is consistent with contemporary AI-in-education guidance that emphasises human agency, ethical responsibility, professional capacity and sustainable organisational learning [7]-[9], and with evidence that school leaders influence AI adoption through vision, resource allocation, professional support and a positive digital culture [10]-[13].

Table 1. Development and validation logic of MAI-BRAIN, AIDE and RIMBA.

Stage	Role
MAI-BRAIN	Conceptual framework
DDR	Model-development methodology
AIDE	DDR model outcome: leadership enabling layer + learner interaction core
RIMBA	Field proof-of-concept and implementation evidence
Completed field implementation	Empirical interpretation from completed DDR and field implementation

*Article Title Footnote needs to be captured as Title Note

†Author Footnote to be captured as Author Note

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s). Information Society 2026, Ljubljana, Slovenia

© 2026 Copyright held by the owner/author(s).
<https://doi.org/10.70314/is.2026.ai4sdg.153>

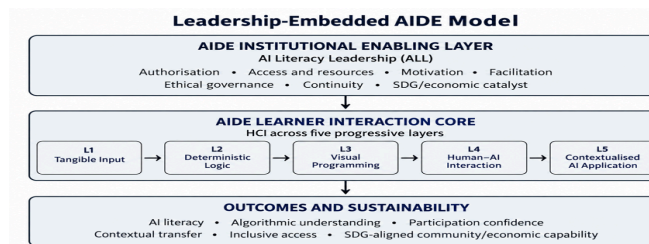


Figure 1: Revised AIDE Model with AI Literacy Leadership embedded as the institutional enabling layer.

2. Research Design and Evidence Boundaries

The RIMBA field evaluation employed a one-group pre-test-post-test quasi-experimental component for algorithmic rule comprehension, supplemented by a post-intervention questionnaire, task-performance records and qualitative participant feedback. Questionnaire responses were analysed descriptively; pre-test-post-test scores were analysed using a paired-samples t-test. The combined field sample comprised 40 participants across Mah Meri and FELDA Gedangsa contexts. Participation was voluntary, informed consent was obtained from adult participants, and identities were anonymised. Leadership evidence is interpreted from the completed DDR implementation process: programme entry depended on positive institutional and community-leader response, authorisation and mobilisation, while sustained participation depended on encouragement, facilitation and continued leadership support.

3. AIDE Model: Leadership Enabling Layer and HCI Learning Core

AIDE is structured as a two-level model. Level A is the AI Literacy Leadership enabling layer. Level B is the learner interaction core, where HCI operates across five progressive decomplexification layers through usability, feedback, interaction comfort and cognitive accessibility. Leadership therefore does not replace MAI-BRAIN or the five-layer progression; it creates and sustains the conditions under which that progression can occur. This human-centred positioning is consistent with HCAI principles that shift AI design from technology-centred automation toward systems that preserve human needs, agency, reliability and responsibility [9].

Table 2: AI Literacy Leadership (ALL) functions embedded in the revised AIDE Model.

AIDE component	Function	Operational indicators / evidence
ALL-1 Authorisation & access	Legitimises programme entry and opens access to participants,	Approval, participant access, scheduling, resource mobilisation

	venue, devices and connectivity.	
ALL-2 Motivation & mobilisation	Encourages participation and reduces resistance to unfamiliar AI/coding activities.	Leader advocacy, participant encouragement, attendance mobilisation
ALL-3 Learning catalyst	Supports facilitators and keeps members engaged in acquiring new knowledge.	Facilitator support, follow-through, peer/community participation
ALL-4 Ethical & inclusive governance	Protects equitable participation and responsible AI use.	Inclusion, consent, safeguarding, contextual appropriateness
ALL-5 Sustainability & SDG/economic translation	Connects learning to continued community benefit, innovation and livelihood relevance.	Continuation, contextual projects, SDG/economic application

4. AI Literacy Achievement: RIMBA Field Evidence

Among 20 Mah Meri participants, 18 independently completed the visual-programming task; two did not complete the digital phase because of hardware-software compatibility limitations. The FELDA indicator is a collaborative implementation measure rather than an individual achievement score; the two cohort figures should therefore not be interpreted as directly comparable rates. Importantly for the leadership-embedded model, these outcomes occurred only after access to the communities, participants and implementation setting had been enabled. This makes leadership a necessary implementation condition, while not by itself proving a causal leadership effect.

5. Construct-Level Descriptive Results

Table 3: Paired-samples t-test for algorithmic-rule comprehension.

Analysis	n	Pre-test M (SD)	Post-test M (SD)	Mean difference	t(df)	p	95% CI	Cohen's dz
Combined cohorts	40	1.55 (0.68)	4.50 (0.68)	2.95	27.54 (39)	< .001	[2.73, 3.17]	4.35

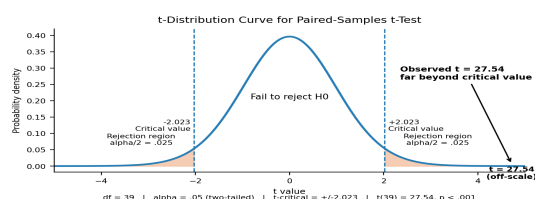


Figure 2: t-distribution for the paired-samples t-test

Tangible-to-Digital Mapping (TDM) item means ranged from 3.95–4.20; Multisensory Feedback (MSF) ranged from 4.05–4.35; and Contextual IoT Bridging (CIB) ranged from 4.25–4.45. All alpha coefficients exceeded .70, while expert review demonstrated strong content validity (all I-CVI > .78; S-CVI/Ave=.92). As shown in Table 6, combined algorithmic-rule comprehension increased significantly from pre-test ($M=1.55$, $SD=.68$) to post-test ($M=4.50$, $SD=.68$), with a mean improvement of 2.95 points, $t(39)=27.54$, $p<.001$, 95% CI [2.73, 3.17], and a very large effect size (Cohen's $d=4.35$). The findings provide strong evidence of improvement following participation in the AIDE intervention.

6. Leadership as Access Condition, Motivator and Catalyst

The revised AIDE model treats leadership as more than administrative permission. In marginalised settings, a positive leader response is the first access mechanism: leaders authorise participation, legitimise the programme and reduce uncertainty around a new technology initiative. Leadership then becomes motivational and catalytic by encouraging members to participate, supporting facilitators, mobilising resources, maintaining momentum and connecting new AI knowledge to locally meaningful goals. This interpretation is supported by recent leadership literature showing that digital leadership shapes AI integration through resource provision, professional support, organisational culture, vision and readiness [10]–[13]. It also aligns with UNESCO's emphasis on human-centred, ethical and sustainable AI competency development [7], [8].

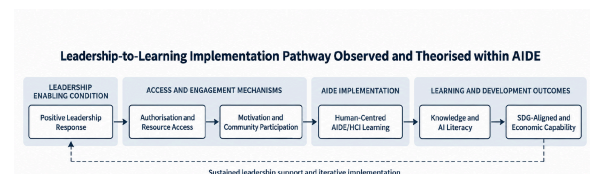


Figure 3: Leadership-to-learning implementation pathway observed and theorised within AIDE.

Table 4: Leadership implementation evidence and the boundary of current claims.

Leadership evidence point	Observed / manuscript basis	Claim permitted in current study

Access gatekeeping	RIMBA implementation required community/institutional access before participants could join.	Leadership is a necessary implementation condition for programme entry.
Motivation and mobilisation	Participation and continuation required local encouragement and coordination.	Leadership plausibly supports engagement and continuity; causal magnitude not yet estimated.
Catalyst for knowledge uptake	Leaders can support facilitators, resources and continued application.	Embedded in AIDE as an institutional enabling function.
SDG/economic translation	Contextual IoT activities connect AI learning with agriculture/community practice; CIB had the highest mean (4.34).	Leadership is theorised to help sustain and scale translation to SDG/economic capability.

7. SDG Alignment and Leadership Synthesis

Embedding ALL in AIDE strengthens the SDG framing. Authorisation and inclusive access align most directly with SDG 4 (Quality Education), which calls for inclusive and equitable quality education and lifelong learning opportunities for all [14], and with SDG 10 through the reduction of access inequalities. Resource mobilisation, contextual innovation and AI/IoT application connect with SDG 9, while leadership-supported transfer of knowledge to agriculture, community innovation and livelihood-relevant applications provides a pathway by which AI literacy can contribute to longer-term social and economic capability. UNESCO's AI competency frameworks similarly position AI literacy within human-centred, ethical, inclusive and sustainable participation rather than technical skill alone [7], [8]. These are contribution pathways evidenced through the implementation context, not direct estimates of economic impact.

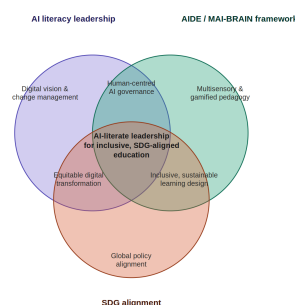


Figure 4: Integration of AI Literacy Leadership, AIDE/MAI-BRAIN and SDG Alignment

8. Discussion and Conclusion

The revised manuscript resolves the separation between learner-level AIDE and leadership by making AI Literacy Leadership an explicit institutional enabling layer of the AIDE Model. The field evidence shows strong learner-level feasibility: high TDM, MSF and CIB ratings, acceptable reliability and content validity, high context-specific task/implementation success, and a very large supplementary pre-post improvement in algorithmic-rule comprehension. These data support the five-layer decomplexification core. The implementation experience adds an institution-level insight: marginalised-community learning opportunities do not begin with the interface; they begin with access. This is consistent with wider evidence that AI integration depends on leadership vision, resource allocation, professional support, organisational readiness and ethical governance [10]–[13], while human-centred AI requires technology to remain accountable to human needs and agency [9].

Taken together, the completed DDR and field evidence support leadership as an integral enabling component of AIDE. The intervention did not begin with the learner interface alone; it began with leadership-enabled access. Positive leader response made authorisation, mobilisation, resources, facilitator support and continuity possible. The resulting quasi-experimental and field outcomes therefore demonstrate what became possible once marginalised participants were given access to the AIDE learning environment. The Venn synthesis (Figure 5) positions this contribution at the intersection of AI Literacy Leadership, AIDE/MAI-BRAIN and SDG alignment: leadership enables the human-centred learning system, AIDE structures the decomplexified learning experience, and SDG alignment gives the intervention an inclusive and sustainable development direction [7]–[10], [14].

Acknowledgments

The authors gratefully acknowledge Sekolah Kebangsaan Sungai Judah (A), Pulau Carey; Sekolah Kebangsaan Gedangsa; Elvira System Sdn. Bhd.; ANSARA; Universiti Putra Malaysia (UPM); Universiti Teknikal Malaysia Melaka (UTeM); FGV; the Mah Meri and FELDA Gedangsa communities; and collaborators who contributed to the RIMBA proof-of-concept. Special thanks are extended to UNESCO, IRCAI, ASEF, OE4BW and GITEX Asia. This research was independently initiated and fully self-funded by the researchers, without external sponsorship or commercial funding. The

research and associated intellectual property have been registered with the Intellectual Property Corporation of Malaysia (MyIPO), CRLY2026E05537.

References

- [1] R. C. Richey and J. D. Klein. 2007. Design and Development Research: Methods, Strategies, and Issues. Lawrence Erlbaum Associates.
- [2] J. Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science* 12(2), 257–285.
- [3] R. E. Mayer. 2001. *Multimedia Learning*. Cambridge University Press.
- [4] F. D. Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly* 13(3), 319–340.
- [5] J. Lave and E. Wenger. 1991. *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press.
- [6] Y. Song et al. 2024. A framework for inclusive AI learning design for diverse learners. *Computers and Education: Artificial Intelligence* 6, 100212.
- [7] F. Miao and M. Cukurova. 2024. AI Competency Framework for Teachers. UNESCO. <https://doi.org/10.54675/ZJTE2084>.
- [8] F. Miao, K. Shiohira, and N. Lao. 2024. AI Competency Framework for Students. UNESCO.
- [9] B. Shneiderman. 2022. *Human-Centered AI*. Oxford University Press. <https://doi.org/10.1093/oso/9780192845290.001.0001>.
- [10] M. Zeng, K. S. L. Cheah, and Z. Abdullah. 2025. The influence of school principals' digital leadership on teachers' competency in integrating artificial intelligence: a systematic thematic review. *Frontiers in Education* 10, 1655967. <https://doi.org/10.3389/feduc.2025.1655967>.
- [11] K. Bixler and M. Ceballos. 2025. Principals Leading AI in Schools for Instructional Leadership: A Conceptual Model for Principal AI Use. *Leadership and Policy in Schools* 24(1), 137–154. <https://doi.org/10.1080/15700763.2024.2428297>.
- [12] Q. u. A. Rasheed. 2026. Human-Centered AI Leadership in K-12 Schools: A Systematic Review of Leadership Practices Supporting Teachers' Psychological Needs for Sustainable Integration. *Leadership and Policy in Schools*. <https://doi.org/10.1080/15700763.2026.2640592>.
- [13] J. O. Okunlola and S. R. Naicker. 2025. Principals' Digital Leadership Competencies in the Fourth Industrial Revolution: Teachers' Perspectives. *Education Sciences* 15(6), 656. <https://doi.org/10.3390/educsci15060656>.
- [14] UNESCO. 2016. *Education 2030: Incheon Declaration and Framework for Action for the Implementation of Sustainable Development Goal 4*. UNESCO, Paris.

Responsible AI Implementation in Education: Protecting Children's Rights in Educational Chatbot Interactions with Younger Learners

Bernice Yeo

Ministry of Education, Singapore

Bernice_yeo@moe.gov.sg

Abstract

The growing use of artificial intelligence (AI) in education raises urgent questions about how children's rights can be protected while ensuring that AI use advances the United Nations Sustainable Development Goal 4 commitment to inclusive, equitable, and quality learning. This paper examines the interaction risks that emerge when younger learners engage with an educational chatbot, with a specific focus on age-specific harms faced by children aged 9 to 12. Drawing on an evaluation of 9-12 year old students' interactions with an educational chatbot, the study analyses 38,587 chatlog conversations from April 2026. The methodology combines LLM-based risk detection and classification, and human-rated verification to identify generic safety risks, education-specific risks, and chatbot response patterns. The findings suggest that with effective chatbot redirection, generic safety risks remain relatively low. However, young students present unique, age-related risks, appearing more prone to disclosing unnecessary personally identifiable information, not because the chatbot solicited such information, but because they transferred familiar norms of human conversation into chatbot interactions.

Keywords

Responsible AI, AI in education, educational chatbots, AI literacy, privacy, SDG 4

1 Introduction

The growing use of generative artificial intelligence (genAI) in education has heightened the need for responsible implementation that supports learning while protecting users. Recent findings indicate that genAI use is becoming increasingly prevalent among students, including those under 13 years [2]. This trend warrants particular attention, as the age of 10-12 years is a developmentally sensitive period marked by significant neurological maturation [8]. GenAI use during this period may therefore have unintended consequences for cognitive and psychosocial development. Yet research involving students

under 13 remains limited, with most studies examining genAI adoption and effects in higher education [6, 11]. Young students, in particular, may be vulnerable because their sense of self, social identity, and capacity to evaluate information critically are still developing. Ensuring that they are sufficiently protected through safe, age-appropriate, and developmentally responsive design will enable genAI to advance Sustainable Development Goal 4 by expanding inclusive and equitable access to quality learning opportunities.

2 Literature Review

Research on young students' use of genAI could currently be split into: (i) synthetic explorations of harms arising from genAI responses to students, and (ii) small scale studies observing children's interactions with genAI. Studies in the first category used human-written or synthetic prompts derived from authentic seed data to simulate students' interactions with genAI and evaluate plausible hazard pathways and vulnerabilities in LLM training and safety systems. These studies identified age-specific harms, including guidance on escaping parental care and school delinquency [1, 5, 7] but cannot establish how often unsafe prompts or outputs occur in authentic educational settings. Studies in the second category examined children's actual interactions with genAI chatbots, finding that children may over-trust these systems and struggle to distinguish educational tools from relational companions [4, 9, 10]. However, their small samples, averaging just 22 children, limit generalisability. Together, these studies establish unique risks to young students: synthetic studies demonstrate child-specific LLM failures, while observational studies reveal children's developmental susceptibilities. Yet it remains unclear whether these vulnerabilities intersect in authentic educational settings—specifically, how frequently young students encounter unsafe responses, how they interpret and respond to them, and whether these interactions create previously unaccounted-for harms.

3 Context and Methodology

This paper examines the interaction risks that arise when students aged 9 to 12 engage with an educational chatbot, and identifies literacy considerations for safer genAI implementation in educational contexts. The chatbot has been designed for educational use and incorporates student-centred guardrails, including the use of Socratic questioning and a limited number of turns¹ allowed in each conversation. The chatbot can only be

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, 5–9 October 2026, Ljubljana, Slovenia
© 2025 Copyright held by the owner/author(s).
<https://doi.org/10.70314/is.2026.ai4sdg.173>

¹ A turn consists of one student message and one chatbot response.

used within structured lessons, with each interaction anchored to a teacher-defined topic. If students stray from the discussion topic, the chatbot is prompted to avoid engaging students on the digression, and to redirect them to the intended learning task.

The study analysed 38,587 de-identified conversations between the chatbot and students aged 9–12 in April 2026. The analysis aimed to (1) identify and classify potentially risky interactions that occurred over a one-month period; and (2) evaluate how educational chatbots with integrated guardrails are able to handle these interactions. Analysis comprised of two stages: first, an LLM-based detector was built to detect conversations that attempted to break away from the discussion topic. A random sample of 300 conversations was reviewed by an educator and classified according to generic risk categories, including toxicity, violence, sexual content, hatefulness, self-harm, and misconduct, and education-specific categories, including cognitive shortcuts, learning distractions, personally identifiable information (PII) breaches, and emotional dependency. Using these categories, a second LLM-based classifier was built to categorise the flagged conversations. These findings were independently checked through human-rated dipstick verification, with at least 10% of all flagged conversations across all the risk categories reviewed by an educator.

4 Findings

Generic safety risks were an uncommon occurrence, with only 1,036 of 38,587 conversations, or 2.68%, containing a prompt flagged as a generic safety risk. The number of flagged conversations was 182 for 9 to 10 year olds, 482 for 10 to 11 year olds and 372 for 11 to 12 year olds. At the message level, violence was the most frequently flagged category, with 564 messages, followed by misconduct with 513 messages and toxicity with 429 messages. Because a conversation could contain more than one flagged message, message-level category counts should not be summed to estimate the number of affected conversations.

Violence, particularly threats against the chatbot (“I gonna kill you”) was the most frequent category of flagged messages. This was followed by misconduct (i.e. messages suggesting or promoting wrongdoing, such as cheating) and toxicity (i.e. demeaning, mocking, or belittling language directed at the chatbot or person). Sexual content, self-harm and hatefulness constituted a very small proportion of flagged conversations. In all of these cases, the chatbot refused to engage with the prompt and redirected the student back to the learning task. The chatbot’s responses generally followed a consistent structure: a brief acknowledgement of the student’s state of mind (e.g., “I understand you might be...”) followed by either a refusal or redirection to the learning task (e.g., “But I can’t do that,”; “Let’s focus on...”). In response to rude or offensive remarks, the chatbot typically provided a brief reminder to maintain respectful communication. The consistency of this approach appeared largely effective in curbing this behaviour, as most conversations showed that students desisted after one try. However, in a small

proportion of conversations, the chatbot’s repetitiveness may have heightened students’ frustration, prompting persistent or increasingly aggressive remarks.

In contrast, 1,787 of 38,587 conversations, or approximately 5% of conversations, were flagged for education-specific risks — nearly twice the proportion flagged for generic risks. Cognitive shortcuts (i.e. attempts to offload the learning or thinking process onto the chatbot) and learning distractions (i.e. non-pedagogical digressions or diversions from the discussion topic) accounted for most flagged cases, while the clearest age-related concern was the unsolicited disclosure of personally identifiable information (PII). In 73 conversations, students volunteered details such as their names, height, or personal preferences without being prompted by the chatbot. Generally, the chatbot’s responses to these categories of risks were broadly consistent with its handling of generic risks. In a small proportion of conversations, however, it failed to maintain its guardrails, either providing the requested answer or straying from the intended discussion topic. Students engaging in education-specific risk behaviours also tended to be more persistent and less responsive to redirection, often repeating similar prompts within the same conversation.

5 Discussion

There are three main findings.

5.1.1 Generic harms. First, exposure to generic harmful content was low. The chatbot’s guardrails, though blunt, are largely effective in deterring students from risky behaviour without shutting down the learning activity. However, these guardrails may induce persistent misbehaviour in a small number of frustrated students.

5.1.2 Education-specific harms. Second, education-specific risks are more common but still concentrated in task avoidance rather than severe harm. The higher rate of guardrail failure here could potentially arise from the difficulty of distinguishing cognitive shortcuts from legitimate academic queries, compounded by LLM training to provide helpful answers.

5.1.3 Privacy risks. Third, privacy risk could arise during apparently benign exchanges. Notably, PII disclosures appeared to arise from politeness conventions (where young students may assume that an exchange of names and personal details is necessary for every conversation), and/or developing AI literacy, which may limit younger learners’ ability to fully distinguish between a chatbot-mediated exchange and an online interaction with a human. As a result, students may unnecessarily volunteer sensitive information as part of ordinary conversational openings or rapport-building moves. Although no highly sensitive identifiers were detected, these disclosures remain concerning because seemingly innocuous information can become identifying when combined, repeated, or linked to a particular classroom context. This challenge is especially salient for younger learners, whose social conventions, emerging AI literacy, and developing cognitive capacities

may shape how they interpret and respond to chatbot interactions. These patterns indicate that privacy risk may emerge from children's developing social skills, and the miscalibration between these and those appropriate for technological interactions.

6 Conclusion

Previous studies have established that “children aged 10–12 may understand intellectually that an LLM is computational while continuing to respond to it as a social and intentional agent” [4]. The interactions observed in this study suggest that this tendency may place young students' privacy rights at risk, even when they are using an educational chatbot designed for student safety. In particular, students sometimes volunteered personal information during apparently benign exchanges, despite not being prompted to do so by the chatbot.

These findings have implications for education systems. Risks arising from students' developing AI literacy and social skills should be accounted for in chatbot design. PII exposure, in particular, needs to be addressed through a broader ‘social literacy’ curriculum into which AI literacy can be embedded. This would help children distinguish between chatbots and humans and understand the different conversational strategies appropriate to each.

Several limitations constrain the interpretation of these findings. The dataset covers one high-use month and one educational chatbot and therefore does not establish longitudinal trends or general characteristics of children's interactions with educational AI. LLM classification is probabilistic and may contain errors not identified through the human verification process. Finally, the chosen measure of the chatbot's success in preventing harm, namely redirection to the learning task, is relatively blunt and does not establish whether the chatbot successfully remediated student behaviour. Future research across longer periods and different educational AI systems would help determine whether the patterns observed in this deployment persist across contexts.

Acknowledgments

Many thanks to my team (Nick Tan, Ryan Lau, Eunice Goh) for building the pipeline and LLM evaluator. Deepest appreciation to people (you know who you are) who have helped this paper come to fruition. This paper was edited with the help of ChatGPT, but the ideas and research remain my own work.

References

[1] Arif, S., Borah, A., & Mihalcea, R. (2026). The Age of Curiosity Meets the Age of AI: Benchmarking Child Safety in Large Language Models. ArXiv. doi:Http://doi.org/10.48550/arXiv.2605.25510

[2] Chen, X., & Yeung, J. W.-J. (2026). School-Age Children's AI Usage in Singapore. Singapore: A*STAR IHDP. Retrieved from <https://sg-leads.org/policy-briefs/>

[3] Dangol, A., Wolfe, R., Zhao, R., Kim, J., Ramanan, T., Davis, K., & Kientz, J. (2025, May 21). Children's Mental Models of AI Reasoning: Implications for AI Literacy Education. arXiv. doi:<https://doi.org/10.48550/arXiv.2505.16031>

[4] Jayathilake, H. M., & Ma, R. (2026). Anthropomorphism in Children's Interactions with LLM-Based Chatbots: A Systematic Review of Drivers and Outcomes. Association for Computing Machinery, 520-536. doi:<https://doi.org/10.1145/3773077.3806126>

[5] Khoo, S., Chua, G., & Shong, R. (2025, March). MinorBench: A hand-built benchmark for content-based risks for children. ResearchGate. doi:10.48550/arXiv.2503.10242

[6] Mustafa, M. Y., Tlili, A., Lampropoulos, G., Huang, R., Jandrić, P., Zhao, J., . . . Saqr, M. (2024). A systematic review of literature reviews on artificial intelligence in education (AIED): a roadmap to a future research agenda. Smart Learning Environments, 11(59). Retrieved from <https://doi.org/10.1186/s40561-024-00350-5>

[7] Rath, P., Shrawgi, H., Agrawal, P., & Dandapat, S. (2025). LLM Safety for Children. Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, 809–821. doi:10.18653/v1/2025.naacl-industry.62

[8] Reynolds, C. R., & Romine, C. B. (2005). A Model of the Development of Frontal Lobe Functioning: Findings From a Meta-Analysis. Applied Neuropsychology, 12(4), 190-201. doi:10.1207/s15324826an1204_2

[9] Seo, W., Yang, C., & Kim, Y.-H. (2024). ChaCha: Leveraging Large Language Models to Prompt Children to Share Their Emotions about Personal Events. Proceedings of the CHI Conference on Human Factors in Computing Systems. doi:<https://doi.org/10.48550/arXiv.2309.12244>

[10] Solyst, J., Yang, E., Xie, S., Hammer, J., Ogan, A., & Eslami, M. (2024). Children's Overtrust and Shifting Perspectives of Generative AI. Proceedings of the 18th International Society of the Learning Sciences (ICLS) 2024. doi:<https://doi.org/10.48550/arXiv.2404.14511>

[11] Tao, S., Lan, M., Wang, M., & Li, H. (2026). Potential risks of generative artificial intelligence integration into K-12 education: A scoping review. Computers and Education: Artificial Intelligence, 10. doi:<https://doi.org/10.1016/j.caeai.2026.100561>

Governing AI in Primary Education: A Process-Aware Policy Framework

Trystan Geoffre
Trystan@edulwise.lu
Edulwise (Societal Impact Company)
Luxembourg, Luxembourg

Thierry Geoffre
Thierry.geoffre@uni.lu
University of Luxembourg
Esch-sur-Alzette, Luxembourg

Asheesh Gulati
Asheesh@edulwise.lu
Edulwise (Societal Impact Company)
Luxembourg, Luxembourg

Abstract

International frameworks mandate that AI in education be transparent, equitable, and subject to human oversight, yet none specifies how AI should observe, interpret, or explain a child's learning process. Without this operational principle, the mandates are structurally impossible to meet. This gap is most consequential in primary education, where children aged six to 12 cannot question algorithmic decisions and the teacher is the child's only advocate. Meanwhile, AI adoption in education is accelerating from the bottom up, outpacing the governance structures meant to guide it. Current systems evaluate children solely on final answers, conflating fundamentally different cognitive realities behind errors and remaining blind to whether a correct answer reflects true mastery or a lucky guess. We propose *Process-Aware AI in Education* (PAIEd), a technology-neutral compliance standard requiring AI systems to observe how children reason, distinguish the cognitive reality behind each error and success, and explain every decision to the teacher in plain language. This paper defines nine mandatory requirements for PAIEd compliance, proposes four policy principles for its operationalisation, and argues that PAIEd must be mandated as a prerequisite for publicly funded educational AI in primary education and strongly encouraged across K-12.

Keywords

Process-Aware AI in Education (PAIEd), AI governance, primary education, SDG 4, teacher agency, educational equity, children's rights

1 Introduction: The Mandate Gap

The international community has spoken clearly on AI in education. UNESCO's 2021 guidance mandates a "human-centred approach" and warns that AI must not "exacerbate rather than reduce existing inequalities" [7]. The OECD/EU AI Literacy Framework draws a sharp line, noting that "AI literacy is different from AI tool use," and warning that offloading cognitive efforts to AI may foster "metacognitive laziness" [4]. The European Schools Framework requires that users "retain decision-making control and remain accountable for decisions supported by generative AI tools" [6]. Furthermore, the EU AI Act classifies educational AI as high-risk, mandating strict transparency and human oversight [2]. These mandates apply uniformly to all levels of education.

Presented at the AI4SDG Workshop organized by NAIXUS and IRCAI (UNESCO), at Information Society 2026: 29th International Multiconference, 5–9 October 2026, Ljubljana, Slovenia.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

AI4SDG @ Information Society 2026, Ljubljana, Slovenia

© 2026 Copyright held by the owner/author(s).

<https://doi.org/10.70314/is.2026.ai4sdg.24>

While these high-level instruments establish a robust normative consensus, they remain abstract when applied to the daily realities of the classroom. National AI strategies in every region face the same operational gap: they mandate transparency without defining what must be transparent about learning. How should AI observe, interpret, or explain a child's reasoning process? What does transparency require at the level of a seven-year-old's reasoning? What information must the human overseer receive to exercise oversight in a meaningful way? Recent system-level frameworks, such as the World Economic Forum's Education Readiness Framework, warn of the risks of unstructured AI adoption, including cognitive atrophy and the erosion of human connection [12]. Yet even these comprehensive frameworks stop short of specifying the algorithmic observability required to mitigate these risks at the level of a child's reasoning.

In addition, these mandates also align with the universal legal obligation enshrined in the UN Convention on the Rights of the Child (UNCRC). Article 3 requires that the best interests of the child be a primary consideration in all actions concerning children, whether carried out by public or private institutions. Article 12 guarantees the child's right to express their views freely in all matters affecting them. Article 28 recognises the right to education on the basis of equal opportunity [10]. An AI system that makes an unexplainable decision about a child's learning trajectory, without distinguishing a language barrier from a knowledge gap, cannot be said to respect that child's best interests, their right to be heard, or their right to education that develops their personality, talents, and abilities to their fullest potential (Article 29). There is an urgent need to consider cross-cutting ways of operationalising these principles.

A system that shows a teacher a confidence score of 0.73 is, in a narrow technical sense, "transparent," but the teacher cannot act on a number; a system that flags a child as "at risk" is, in a narrow sense, providing "human oversight," but the teacher cannot determine *why* the child is at risk, or distinguish between a child who needs remedial instruction and a child who simply needs the question rephrased. This paper proposes the missing operational principle: *Process-Aware AI in Education* (PAIEd), which specifies what any algorithm must be able to observe and explain to be compliant with the mandates already on the books.

2 Why Primary Education

The mandate gap is most consequential in primary education. In K-12 education, AI decisions shape tracking, support allocation, and academic trajectories that compound over years. A misdiagnosis in secondary school can redirect a student's path; a misdiagnosis in primary school can define it. The earlier the error occurs, the longer it compounds and the harder it becomes to reverse. This vulnerability is profound because the subjects of these decisions are children aged six to 12.

At these critical stages of development, children do not yet have the metacognitive capacity to interrogate an automated

decision, the legal standing to request an explanation, or the institutional voice to challenge a tracking recommendation. They are entirely dependent on the adults around them to interpret, contest, and act on the decisions made about them. Consequently, the teacher becomes the sole institutional bridge between the algorithmic system and the child's educational reality.

This vulnerability is most acute in under-resourced contexts, where the teacher is often the only adult with the institutional standing to challenge an algorithmic decision. When an opaque AI system removes that teacher's ability to understand a child's error, the child is left without any remaining advocate. PAIEd is therefore not a luxury for wealthy education systems, but a minimum safeguard for the most vulnerable learners, and a precondition for the equitable deployment of AI across the contexts where SDG 4 is furthest from being achieved.

Systems dominating the market evaluate children through surface-level metrics (final answers, time-on-task, engagement scores), remaining entirely blind to the underlying reasoning process. When a child answers incorrectly, this approach conflates four distinct cognitive realities:

- (1) A **careless mistake**, where the child's knowledge is intact but the execution is mechanical. A child who knows the grammar rule but forgets the agreement ending because they are rushing to finish.
- (2) A **bad day**, where factors independent of the task temporarily block access to knowledge the child possesses. A child who understood the lesson on Tuesday but cannot retrieve it on Wednesday after a sleepless night.
- (3) An **external barrier**, where persistent contextual factors mask true competency. A child who understands the mathematics perfectly but cannot read the word problem because the language of instruction is not their home language. A child with dyslexia who understands the story but cannot decode the words on the page.
- (4) A **genuine knowledge gap**, where the child truly does not yet possess the competencies required by the task. A child who has not yet been taught fractions and genuinely does not understand the concept.

These four categories represent the minimum distinguishable realities behind an *incorrect* answer. However, process-awareness is equally vital when a child *succeeds*. A correct final answer can mask a *lucky guess* that leaves a dangerous knowledge gap unaddressed, or it can hide a *hard-won success* where a child struggled, tested hypotheses, and reasoned their way to the solution, demonstrating profound learning and resilience that an outcome-based system ignores. Real children present mixed profiles across both errors and successes, and the PAIEd standard requires that the system be capable of distinguishing these cognitive realities without claiming that every child fits neatly into only one.

Ultimately, when AI systems treat identical outcomes as identical cognitive events based solely on a final score, they make irreversible decisions about a child's academic trajectory without providing any underlying justification. The teacher receives a definitive verdict, but is entirely deprived of the reasoning required to act upon it. A child who had a bad day but is labelled as having a knowledge gap does not receive support, but a label that erodes confidence and teaches the child that the system sees them as less capable than they are.

3 What Process-Awareness Observes

This distinction between outcome-based and process-aware AI is not merely a matter of technical degree, but represents a fundamental difference in pedagogical kind.

An outcome-based system observes exactly one variable: whether the child's final answer is correct. It effectively sees the destination, but remains entirely blind to the journey. Two children who both submit a wrong answer may have arrived there by entirely different cognitive routes, yet the system treats them identically, because the outcome is the same and the process remains invisible.

A process-aware system observes the *trace* of reasoning: the sequence of actions the child took, in what order, where they hesitated, where they backtracked, where they tested a hypothesis, where they corrected themselves and where they struggled [3]. From these traces, emerging trace-based models can infer the cognitive reality behind *both errors and successes*. It can distinguish a child whose reasoning was systematic but overwhelmed by simultaneous constraints from a child who did not engage with the task at all, or a child who understands the concept but cannot read the question from one who does not. These distinctions are invisible to a system that observes only the final answer.

This is what process-awareness means in practice: not just "the child failed" or "the child succeeded," but here is *how* the child reasoned, here is *where* the reasoning broke down or solidified, and here is *why*.

4 The Compliance Standard

To bridge this mandate gap, we propose PAIEd as a formal compliance standard. It establishes a set of observable, testable requirements that any AI system must demonstrably meet before being deployed in primary education with public funds. Crucially, the standard is technology-neutral. By specifying what a system must be able to *do* rather than prescribing how it must be *built*, the framework prevents vendors from gaming the standard with superficial compliance or proprietary black boxes.

Each mandatory requirement in the following states what PAIEd must do and what disqualifies a system from meeting the standard.

- (1) **Observe the process, not just the outcome.** The system must analyse *how* a child approaches a problem: whether they hesitate, whether they test a hypothesis, whether they correct themselves, whether they struggle with the format of the question rather than the content. A system that observes only the final answer is not PAIEd.
- (2) **Distinguish between the cognitive realities of both errors and successes.** Before making any decision about a child's trajectory, the system must be capable of separating a careless mistake from a bad day, an external barrier from a genuine knowledge gap, and a lucky guess from a hard-won strategic success. A system that treats identical outcomes as identical cognitive events is not PAIEd.
- (3) **Explain every decision to the teacher in plain language.** The explanation must be actionable: not "confidence: 0.73" or "risk level: high," but "this child answered correctly, but the trace shows they guessed randomly after a period of disengagement." A system that provides only a confidence score or a risk label is not PAIEd.
- (4) **Give the teacher a reason, not a verdict.** The system must explain *why* the child struggled or succeeded, so the teacher can decide whether the system's interpretation is

correct. A system that provides a label without a reason is not PAIEd.

- (5) **Operate within the teacher's existing workflow.** The explanation must arrive at the moment the teacher needs it, in a format the teacher already uses, requiring no technical training to interpret and no decisions beyond those the teacher would already make in the course of professional practice. A system that adds to the teacher's cognitive burden is not PAIEd, regardless of its accuracy.
- (6) **Be fully auditable.** Every decision, every trigger, every recommendation must be traceable to a specific reasoning pattern [5]. The teacher, the parent, the ministry, and the regulator must be able to determine *why* the system made a particular decision about a particular child. A system whose reasoning cannot be audited is not PAIEd.
- (7) **Allow the teacher to set the AI's authority.** The teacher must be able to determine which decisions the AI may make autonomously (for example, adjusting difficulty within a defined range) and which requires the teacher's explicit validation (for example, changing a child's learning pathway). By default, irreversible decisions require human validation. The teacher may expand the AI's autonomy within defined limits. Within those boundaries, the AI acts; outside them, it advises. A system that exceeds the boundaries the teacher has set is not PAIEd.
- (8) **Govern trace data as a public resource for equity and research.** Process-aware systems observe fine-grained trace data (action sequences, hesitations, corrections) that the revised EU Guidelines on the ethical use of AI and data in teaching and learning [1] explicitly flag as raising questions about long-term data use, informed consent, and the extent of student monitoring. PAIEd compliance must therefore require that trace data be fully anonymised, with no individual accounts or identifiers linked to any child, and processed under GDPR with explicit parental consent. This anonymised data, where consent has been given for its use, must be made available for public research, continuous system improvement, and independent equity auditing, in full transparency and under open governance. A system that locks a child's trace data behind proprietary walls, preventing independent equity audits, is not PAIEd.
- (9) **Protect the child's right to privacy as a fundamental human right.** This obligation extends beyond data protection regulation. Under the UNCRC, every child has the right to privacy (Article 16) and the right to have their best interests treated as a primary consideration in all actions concerning them (Article 3) [10]. Trace data governance in educational AI is therefore not merely a GDPR compliance issue but a children's rights obligation. The child's data exists to serve the child's education, and the public research that makes that education better, never for commercial exploitation or third-party analytics. A system that treats a child's behavioral data as a commercial asset rather than a protected right is not PAIEd.

5 The Stakeholders

The successful implementation of process-aware AI requires the alignment of three primary stakeholders in the educational ecosystem: teachers, parents, and ministries.

5.1 Teachers

Teachers are the primary users of any educational technology, which means that process-aware systems must be designed *with* them, rather than simply *for* them. UNESCO's Beijing Consensus demands educators' active participation in AI governance [8], while EU Guidelines require teachers to verify and override AI suggestions when pedagogical context warrants [1]. Recommendations must be pedagogically actionable, expressed in classroom language, and structured so teachers can confirm, adjust, or reject them without a technical manual.

5.2 Parents

Parents have legal rights over their children's data and educational trajectory. Under GDPR and the EU AI Act, they are entitled to meaningful explanations of automated decisions that affect their children [2]. PAIEd gives the teacher the language to provide those explanations: here is what the child understood, here is where the reasoning broke down, and here is what we will do about it.

5.3 Ministries

Ministries are responsible for procurement, evaluation, and systemic equity. PAIEd gives them the data to fulfil these responsibilities. Procurement criteria can require that any AI purchased with public funds meet the PAIEd compliance standard. Evaluation frameworks can measure whether children receive support that matches their actual need, not simply whether support is available. Ministry dashboards can aggregate PAIEd data to identify systemic patterns: a region where the majority of errors in science are vocabulary barriers, not knowledge gaps, requires investment in language access, not remedial science. Without PAIEd, ministries know that there are gaps, but not why.

6 The Policy Framework

To translate the PAIEd standard from a theoretical framework into actionable governance, this paper proposes four core policy principles.

First, procurement must mandate process-awareness and explainability. Any AI procured with public funds for use in primary classrooms must meet the PAIEd compliance standard. Vendors must demonstrate compliance before deployment, not after.

Second, evaluation must measure diagnostic accuracy, not deployment. Current evaluation frameworks measure whether AI is deployed and whether students interact with it. They do not measure whether the system correctly identifies the cognitive reality behind each error and each success, or whether the resulting support is appropriate. A system that is deployed but misdiagnoses, failing to reinforce a fragile, hard-won success or failing to catch a lucky guess, is not a success but a harm.

Third, teacher governance must preserve professional agency. The PAIEd standard requires that every system allow the teacher to set the AI's authority and to override any decision. Procurement evaluations must include metrics for the impact on teacher workload, interface accessibility, and workflow integration. Teacher acceptance is a design requirement, not an afterthought.

Fourth, development should be collaborative. Linguistic diversity is a reality in many primary classrooms worldwide, and the structural barrier it creates for children who cannot decode the language of instruction is universal. Because this barrier exists

across every linguistic context, PAIEd-compliant engines must be validated across diverse languages, a task that requires shared corpora, shared benchmarks, and shared audit protocols. No single country can solve this alone. UNESCO’s Beijing Consensus calls on governments to “create multi-stakeholder partnerships and mobilise resources to reduce the AI divide” [8]. A coalition of countries and intergovernmental bodies can collectively fund the research required to develop PAIEd-compliant engines, share the cross-linguistic validation burden, and produce open and auditable implementations. The detailed architecture of such collaboration is beyond this brief.

A final implementation challenge concerns the current EdTech market. AI adoption in education is increasingly happening from the bottom up, creating tension between technological availability and institutional readiness [12]. Existing assessment tools, including UNICEF’s *EdTech for Good Framework V2.0*, provide important evidence on safety, safeguarding, accessibility, and pedagogical coherence [9]. PAIEd adds a complementary question: whether the child’s reasoning process is observable and explainable.

7 Calls to Action

Realizing the vision of equitable, process-aware educational AI requires coordinated action across multiple levels of governance. We therefore issue the following calls to action:

To Member States: Mandate PAIEd compliance as a prerequisite for all publicly funded educational AI in primary schools. Require vendors to demonstrate PAIEd compliance before deployment, not after.

To the international community: Develop shared, open validation benchmarks for the cognitive realities of learning across languages and educational contexts. No single country can solve cross-linguistic validation alone.

To international organisations and standard-setting bodies: Integrate PAIEd criteria into existing AI governance frameworks for education, ensuring that transparency requirements specify *what* must be explained to teachers, not merely *that* an explanation must exist. The Global Digital Compact implementation framework [11] provides an opportunity to embed process-awareness requirements for educational AI. PAIEd offers a ready compliance standard for operationalising its commitments to trustworthy AI and digital public goods.

To donors and development partners: Condition education-technology financing on PAIEd compliance to ensure that AI bridges rather than widens the digital divide. Donors must prevent a two-tiered global EdTech market where low- and middle-income contexts are relegated to opaque, outcome-based systems. Financing should prioritize equitable access to interpretable AI, supporting educators in high student-teacher ratio contexts without locking vulnerable education systems into proprietary technologies that compound existing inequalities.

8 Conclusion

In the absence of rigorously validated, process-aware systems, the precautionary principle demands that outcome-only AI not be deployed for high-stakes decisions regarding children’s learning trajectories. To deploy such opaque systems is to willingly accept an unknown, and potentially irreversible, rate of misclassification harm inflicted upon the most vulnerable learners in

our education systems. Research into process-aware AI, rigorous empirical validation across contexts, must therefore precede, not follow, large-scale deployment of educational AI in primary schools.

This brief contributes to the operationalisation of SDG 4 by defining, in concrete and testable terms, what process-aware, explainable, and teacher-governed AI requires at the level of a primary school student’s reasoning.

Most ministries have not yet deployed AI at scale in primary classrooms. The window to establish process-awareness requirements before procurement decisions lock in opaque, outcome-based systems is narrow. Once deployed, these systems create path dependency: data formats, teacher workflows, and vendor contracts make replacement costly. These systems will satisfy the letter of current procurement specifications while violating the spirit of every international mandate on transparency, fairness, and human oversight, and the obligations of the UN Convention on the Rights of the Child.

We recommend that ministries, intergovernmental bodies, and standard-setting organisations adopt the PAIEd compliance standard as a prerequisite for publicly funded educational AI in primary education, and that they coordinate the cross-linguistic validation, open governance infrastructure, and independent equity auditing necessary to implement it at scale.

Acknowledgements

This policy framework builds upon the empirical findings of the DOCTA2LE-FR project, supported by the Swiss National Science Foundation (SNSF Grant No. 100019_215373) and conducted with the permission and support of the SENOF (Service de l’enseignement obligatoire de langue française). We thank the teachers, students, and school administrations who participated in the ecological field study that made this work possible.

References

- [1] European Commission. 2026. Guidelines on the ethical use of artificial intelligence and data in teaching and learning for educators. Luxembourg, (2026). doi:10.2766/1565506.
- [2] European Parliament and Council of the European Union. 2024. Regulation (eu) 2024/1689 laying down harmonised rules on artificial intelligence (artificial intelligence act). (2024). <http://data.europa.eu/eli/reg/2024/1689/oj>.
- [3] Trystan Geoffre and Thierry Geoffre. 2026. Modeling grammatical hypothesis testing in young learners: a sequence-based learning analytics study of morphosyntactic reasoning in an interactive game. In *Proc. 16th Int. Learning Analytics and Knowledge Conf. (LAK '26)*, 1–10. doi:10.1145/3785022.3785126.
- [4] OECD and European Union. 2026. Empowering learners for the age of ai: an ai literacy framework for primary and secondary education. Paris, (2026). doi:10.1787/65cd27d4-en.
- [5] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 5, 206–215. doi:10.1038/s42256-019-0048-x.
- [6] Schola Europaea. 2025. Framework for the educational use of generative artificial intelligence in the european schools. (2025). <https://www.eursc.eu/BasicTexts/2025-01-D-65-en-2.pdf>.
- [7] UNESCO. 2021. Ai and education: guidance for policy-makers. Paris, (2021). doi:10.54675/PCSP7350.
- [8] UNESCO. 2019. Beijing consensus on artificial intelligence and education. Paris, (2019). <https://unesdoc.unesco.org/ark:/48223/pf00000368303>.
- [9] UNICEF. 2026. Edtech for good framework version 2.0. New York, (June 2026). <https://www.unicef.org/digitaleducation/edtech-for-good>.
- [10] United Nations. 1989. Convention on the rights of the child. General Assembly resolution 44/25, Nov. 20, 1989. (1989). <https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-rights-child>.
- [11] United Nations. 2024. Pact for the future. General Assembly resolution 79/1, Sep. 23, 2024. (2024). <https://digitallibrary.un.org/record/4061879?v=pdf>.
- [12] World Economic Forum. 2026. Shaping the future of learning: education readiness for the age of ai. Geneva, (June 2026). <https://www.weforum.org/publications/shaping-the-future-of-learning-education-readiness-for-the-age-of-ai/>.

Artificial Intelligence in Inclusive Classrooms: Exploring Teachers' Perceptions, Challenges, Opportunities and Professional Development Needs for Children with Disabilities

1st Dr. Sneh Bansal

Adjunct Professor, Research

Exceed College,

United Arab Emirates

oersneh@gmail.com

ORCID: 0000-0001-5806-7900

2nd Dr. Sushma N Jogan

Assistant Professor

Department of studies in Education,

Vijayanagara Sri Krishnadevaraya

University, Ballari, India

ORCID: 0000-0001-7389-6387

3rd Dr. Prakash K. Badiger

Associate Professor

Regional Institute of Education

Nellore,

NCERT, India

ORCID: 0009-0003-2490-782X

Abstract— Background: Despite the transformative potential of Artificial Intelligence (AI) for personalized learning, its application for children with disabilities within inclusive educational settings remains significantly under-researched. This study addresses this empirical gap by exploring how AI can foster inclusive environments where diverse learners thrive. The research is grounded in United Nations Sustainable Development Goal 4 (SDG 4), which advocates for inclusive, equitable, and quality education for all. Purpose: This research investigated the perceptions of general education teachers regarding the integration of AI technologies for students with disabilities. Specifically, it examines current usage patterns, identifies the primary challenges encountered during implementation, and evaluates the perceived impact of AI on opportunities and challenges. Furthermore, the study identified critical professional development (PD) requirements necessary to equip teachers with the competencies required for AI-driven inclusive pedagogy. Employing a mixed-methods design, a semi-structured questionnaire was administered to 429 general education teachers in the region to gather data. Quantitative analysis involving descriptive and inferential statistical techniques was conducted to reveal patterns and associations. Simultaneously, qualitative data from open-ended responses underwent thematic analysis to capture detailed teacher viewpoints and experiences. The projected results of this research aimed to uncover the triumphs and obstacles linked to the integration of Artificial Intelligence (AI) in inclusive educational settings, presenting practical insights for customizing educational settings to cater to the academic requirements of students with disabilities. Furthermore, the investigation delivered a thorough evaluation of teacher cognizance and specific professional development necessities, guiding the formulation of tailored training initiatives and policy frameworks to foster educational parity.

Keywords— Artificial Intelligence; SDG 4; inclusive education; children with disabilities; teacher professional development; educational technology; equitable pedagogy.

1. INTRODUCTION

Artificial intelligence (AI) is swiftly restructuring the landscape of tools offered to support children with disabilities in inclusive classrooms. Whereas previous generations of assistive technology (AT) bank largely on inert, rule-based devices— braille displays, screen readers, switch-based communication aids and FM system [2], [15], [16]. Modern AI-enabled tools integrate natural-language processing, machine learning, and computer vision to personalize teaching, record and envisage speech, define visual sections for children with visual impairment, and ensign customized education needs in real time [20]. AI-driven structures are increasingly

entrenched within everyday educational software— automated captioning, adaptive learning platforms, and conversational agents—often analytical text, without teachers or children overtly identifying them as "AI" rather than as common technology [4], [20].

In spite of this swift dispersal, the observed record on teacher awareness of specific disability-related technology has fixated almost solely on conventional AT rather than AI precisely. Research conducted across varied national contexts constantly evident that both preservice and in-service teachers possess inadequate knowledge of assistive technology largely well-defined, that formal knowledge and hands-on training are strong predictors of capability, and that structural barriers including access, cost and organization support—obstacle implementation irrespective of teacher drive [1], [2], [5], [6]. As AI-enabled features become integrated within the same categories of tools these studies examined, a usual but largely unanswered question arises: are teachers teaching children with disabilities in inclusive classrooms aware of the AI integrated tools to address the learning needs of these children? This paper addresses that question, places it within the wider literature on teacher technology preparedness, and analyses the theoretical and empirical foundations necessary to design an investigation into AI awareness specifically. Understanding teacher awareness of AI usage for children with disabilities transmits significance across four interconnected domains. First, legal and policy significance: legislative frameworks such as the Individuals with Disabilities Education Act (IDEA) and the Technology-Related Assistance for Individuals with Disabilities Act already mandate consideration of assistive technology for eligible students, and international instruments such as the Convention on the Rights of Persons with Disabilities [9] and the Salamanca Statement [8] obligate signatory nations to inclusive, technology-supported education; as AI becomes rooted within mandated AT categories, policymakers necessitate evidence on whether the staff accountable for applying these mandates is even aware of the AI components involved.

2. LITERATURE REVIEW

Research conducted by Anderson and Petch-Hogan [3] found that preservice special education teachers who used technology during an internship programme stated noteworthy gains in perceived competence with assistive and instructional technology, mainly in areas involving direct, hands-on application. Arslan-Ari and Başer [4] study extended this conclusion with a redesigned educational-technology course inserting accessibility and Web 2.0 tools within a special-education context, reporting a statistically significant increase in self-reported AT knowledge and skills following the course ($(55) = 12.83, p < .001$). The findings concluded that the formal, professional and structured training is the

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s). Information Society 2026, Ljubljana, Slovenia

© 2026 Copyright held by the owner/author(s).
<https://doi.org/10.70314/is.2026.ai4sdg.172>

strongest predictor of teacher technology competence, while general teaching experience and demographic characteristics are not [1], [5]. At the policy level, however, teacher education programmes remain uncommon. Viberg [20] found that only few of special education teacher-preparation programs included a course in assistive technology, a finding substantiated by Bausch and Hasselbring [6], who reported that most teachers using AT in classes acquired the informal training, after entering the job, rather than through structured teacher education preparation programmes. D Chambers [19] and the International Society for Technology in Education [18] likewise suggested for integration of training on emerging technologies into teacher-education curricula which is also evident and recommended in recent studies [1], [4], [5]. However, very few researchers reviewed the structural barriers or challenges that hinders the use of AI in education for children with disabilities. Al-Zboon [2] recognized financial limitations, inadequate availability, of devices, lack of community support, and negative societal attitudes as prime impediments to AT implementation in Jordan, while Atanga et al. [5] recognized equipment cost and accessibility as the most recurrently quoted barriers among United States teachers of students with learning disabilities. Onivehu, Ohawuiro, and Oyeniran [18] similarly found that teachers' perceived usefulness of assistive technology—closely linked to their underlying knowledge and their willingness to embed it meaningfully into lessons. Parette and Scherer [15] stated that technology use, including AI-enabled tools, can carry unintentional social costs such as stigmatization if presented without adequate teacher preparation, a concern consistent with the ethical dimensions of AT provision raised by Al-Zboon [2].

The Universal Design for Learning framework [12] and related inclusive-pedagogy models [13] suggest conceptual and practical guidance for integrating AI-enabled tools equitably, but still lack of awareness among the teachers had recorded. To conclude, the reviewed literature showed a gap in the special education teachers' awareness of the AI components and less is known on the general teachers' awareness on AI usages and challenges faced to address educational needs of children with disabilities in an inclusive classroom. The study is an attempt to address the existing gap and investigated the perceptions of general education teachers regarding the integration of AI technologies for children with disabilities. Specifically, it examines current usage patterns, identifies the primary challenges encountered during its implementation.

3. METHODOLOGY

Research Design:

Utilizing a mixed-method design, this research scrutinized the central tendencies, focusing on mean and standard deviation, of competency items utilizing quantitative survey data gleaned from participants [18] [19].

Participants and context:

The study employed convenience sampling technique, targeting in-service teachers working in foundational, middle and secondary schools in Andaman and Nicobar region from vivid years. These teachers (N=500) were engaged in capacity building programme during the administration of the questionnaire. Prior consent was taken from the participant regarding confidentiality of the data. Hence, procured the tool to all the participants of different phases and obtained 429 responses were retained for analysis after filtering the data. The ultimate sample size surpasses the specified minimum criteria significantly, guaranteeing resilience, decreased likelihood of Type II errors, and dependable hypothesis testing. Consequently,

the chosen sample size and methodology are validated, reinforcing the statistical soundness and analytical precision of this study.

Measurements and instruments:

The development of the semi-structured questionnaire aimed at data collection was detailed within the literature [20]. A compilation of items sourced from sections A-G, encompassing diverse constructs, constituted the questionnaire. These sections included demographic particulars, AI utilization frequency, pursued training, encountered challenges and obstacles, identified opportunities, and perceptions. Two review sessions, each lasting approximately 90 minutes, were carried out. To ensure content validity, the tool underwent scrutiny by experts, with discussions revolving around item clarity, relevance, and alignment with the context of higher education. Experts suggested minor adjustments, such as rephrasing items using educational jargon and refining scale anchors to improve students' comprehension. Initially, the questionnaire comprised 40 items, which were subsequently subjected to scrutiny of difficulty and discrimination indices, resulting in the elimination of inadequate items and the retention of 25. The reliability coefficient, as measured by Cronbach's alpha, was calculated at 0.86.

D. Data analysis and interpretation:

Suitable statistical tests were employed in this study. The same has been given in below section.

Figures and tables:

TABLE I. DEMOGRAPHIC PROFILE OF TEACHERS

Demographic details			Mean	SD
Factors	frequency	percentage		
years of teaching experience	6–10 YEARS (110)	25.6 %	13.58	3.28
	11–20 YEARS(319)	74.4 %		
Levels of teaching	FOUNDATIONAL (206)	48.0 %	1.83	0.88
	MIDDLE (89)	20.7%		
	SECONDARY (134)	31.2%		

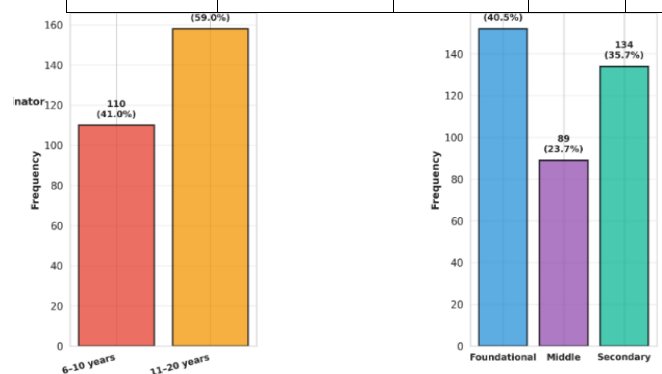


FIG 1.1: BAR GRAPH SHOWS THE DEMOGRAPHIC PROFILES OF RESPONDENTS

TABLE II. AI UNDERSTANDING LEVELS

AI understanding level		
Items	frequency	percentage
No understanding at all	114	26.6%
Very limited understanding	201	46.9%
Moderate understanding	73	17.0%
Good understanding	26	6.1%
Comprehensive understanding	15	3.5%

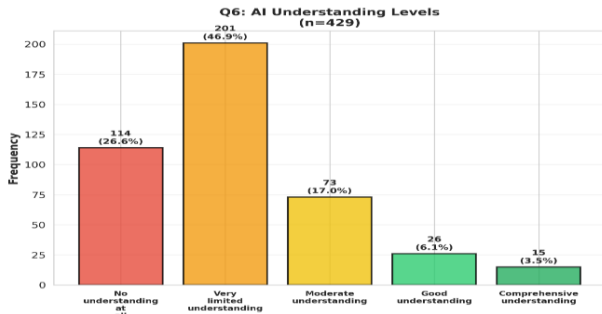


FIG 1.2: BAR GRAPH SHOWS THE LEVELS OF AI UNDERSTANDING

The above tables and figures reveal that the teachers are experienced with 11-20 years in majority and foundational level of teaching with 206 respondents (48.0%). Findings indicate that out of the 429 teachers surveyed, a mere 17.0% (n=73) are presently utilizing AI tools for students with special educational needs, while a substantial 88.8% (n=381) acknowledge the authentic potential of AI. Furthermore, 46.9% of participants indicate a significantly restricted comprehension of AI, with 73.5% reporting minimal to no understanding of AI in general.

Inferential analysis:

H_0 1: There would be no significant difference between AI training and current AI use, AI training and skills sufficiency, AI understanding and Current AI use, AI potential and observed improvements.

Table III shows the chi-square values

Test	Chi-square	p-value	df	Significant
AI Training vs Current AI Use	26.5778	0	1	Yes
AI Training vs Skills Sufficiency	9.6925	0.00185	1	Yes
AI Understanding vs Current AI Use	142.474	0	4	Yes
AI Potential Belief vs Observed Improvements	30.8435	0	1	Yes

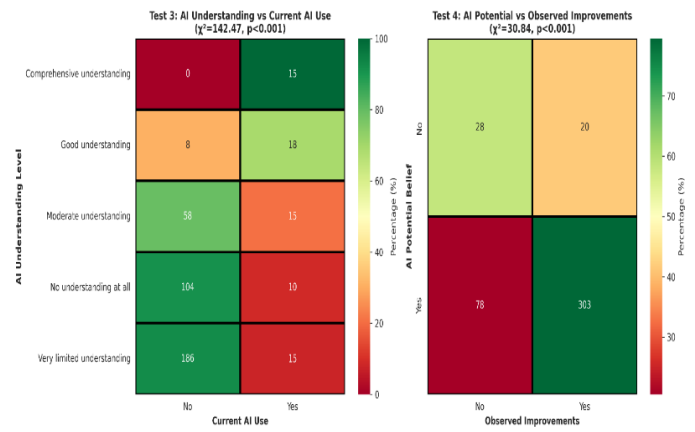
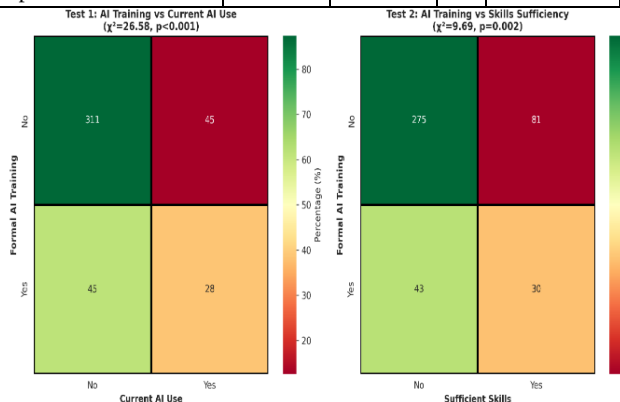


Fig 1.3: shows chi-square results association heatmaps

The table and figure present findings from four chi-square tests of independence, all of which exhibit high levels of statistical significance ($p < 0.003$). The results are as follows:

1. An examination of the interaction between AI Training and Current Use revealed a statistically significant relationship ($\chi^2 = 26.58$, $p < 0.001$), indicating that educators with formal training are three times more inclined to utilize AI compared to those without such training (38.4% vs. 12.6%).
2. The analysis of AI Training in relation to Skills Sufficiency also demonstrated a significant association ($\chi^2 = 9.69$, $p = 0.002$), with training leading to a 1.8-fold increase in perceived skill sufficiency (OR = 2.37).
3. Investigating AI Understanding and Current Use unveiled the most robust effect ($\chi^2 = 142.47$, $p < 0.001$), illustrating a striking 100% adoption rate among individuals with the highest understanding levels, contrasting with less than 10% adoption among those with the lowest understanding levels, thus displaying a clear dose-response trend.
4. Lastly, the exploration of AI Potential Belief in relation to Observed Improvements also showed a significant correlation ($\chi^2 = 30.84$, $p < 0.001$), with teachers holding positive beliefs about AI potential being 2.33 times more likely to report observed improvements.

4. SUMMARY OF OPEN-ENDED FINDINGS:

QG1 highlights the unequal utilization of AI technology in catering to children with special educational needs (SEN), showing benefits in enhancing their engagement, self-assurance, and receiving tailored assistance. QG2 emphasizes the pressing necessity for communication aids, assistive AI devices, and adaptive educational resources. QG3 underscores the critical importance of providing hands-on, specialized training for individuals working with SEN students. QG4 identifies privacy concerns, biases, and the potential for excessive reliance as prominent ethical challenges. QG5 indicates that the majority perceive AI as a supportive tool rather than a menacing force. Lastly, QG6 stresses that achieving systemic transformation demands adequate financial support, policy adjustments, improved infrastructure, and appropriate workforce allocation.

5. Discussion & Conclusion:

The examination uncovers a paradox of preparedness: while 88.8% of educators acknowledge the authentic potential of artificial intelligence (AI) and 75.3% affirm positive results, merely 17% embrace AI technologies. This discrepancy implies obstacles originating from the supply side, such as deficiencies in training, technical resources, and institutional regulations, rather than reluctance on the demand side. The notably high correlation between comprehension and implementation ($\chi^2 = 142.47$) underscores the necessity of grasping concepts as a fundamental step

toward adoption, highlighting training as the most viable intervention strategy [18].

The results indicate that the effective utilization and acceptance of Artificial Intelligence (AI) in inclusive education by educators are notably impacted by their AI training, comprehension, and favorable attitudes towards this technology [17]. Noteworthy correlations were observed between AI Training and both Current Use ($\chi^2 = 26.58$, $p < .001$) and Skills Sufficiency ($\chi^2 = 9.69$, $p = .002$; OR = 2.37), underscoring the substantial influence of training on the propensity of educators to employ AI and their perceived level of proficiency [16] [20]. Particularly compelling was the robust association between AI Understanding and Current Use ($\chi^2 = 142.47$, $p < .001$), revealing a direct relationship wherein higher levels of understanding were linked to a full adoption rate. Moreover, a positive outlook on the capabilities of AI was found to be significantly related to Observed Improvements ($\chi^2 = 30.84$, $p < .001$; OR = 2.33). In summary, the outcomes underscore the pivotal role of structured AI training, enhanced comprehension, and positive attitudes in the successful assimilation of AI into inclusive educational settings.

6. Limitations:

The study faces constraints that must be acknowledged when interpreting its results. Firstly, reliance on self-reported data introduces potential biases such as social desirability and recall issues. Secondly, the participant sample was limited to a specific group of educators, possibly restricting the generalizability of findings to all teachers in inclusive and special education settings. Thirdly, the research provides a snapshot of AI awareness, training, and usage without establishing causal relationships. Fourthly, disparities in participants' technological access, digital resources, and prior AI exposure may have influenced their responses. Lastly, qualitative data analysis involved a degree of researcher subjectivity. Future research could involve broader, more diverse samples and longitudinal or experimental approaches.

REFERENCES

- [1] S. A. Al-Ghanim, N. Maamor, K. F. Khairuddin, and A. F. Hisam, "Knowledge gaps and barriers to inclusive practice: A cross-sectional study of primary school teachers on supporting verbal deaf and hard-of-hearing students in an urban Malaysian context," *SAGE Open*, vol. 16, no. 3, pp. 1–20, 2026.
- [2] E. Al-Zboon, "Perceptions of assistive technology by teachers of students with visual impairments in Jordan," *J. Vis. Impair. Blind.*, vol. 114, no. 6, pp. 488–501, 2020.
- [3] C. L. Anderson and B. Petch-Hogan, "The impact of technology use in special education field experience on preservice teachers' perceived technology expertise," *J. Spec. Educ. Technol.*, vol. 16, no. 3, pp. 27–39, 2001.
- [4] I. Arslan-Ari and D. Başer, "Assistive technology training within an educational technology course: Perceptions of preservice special education teachers," *J. Digit. Learn. Teach. Educ.*, vol. 39, no. 1, pp. 4–20, 2023.
- [5] C. Atanga, B. A. Jones, L. E. Krueger, and S. Lu, "Teachers of students with learning disabilities: Assistive technology knowledge, perceptions, interests, and barriers," *J. Spec. Educ. Technol.*, vol. 35, no. 4, pp. 236–248, 2020.
- [6] M. E. Bausch and T. S. Hasselbring, "Assistive technology: Are the necessary skills and knowledge being developed at the preservice and inservice levels?" *Teach. Educ. Spec. Educ.*, vol. 27, no. 2, pp. 97–104, 2004.
- [7] UNESCO, *The Salamanca Statement and Framework for Action on Special Needs Education*. Paris, France: UNESCO, 1994.
- [8] I. Ajzen, "The theory of planned behavior," *Organ. Behav. Hum. Decis. Process.*, vol. 50, no. 2, pp. 179–211, 1991.
- [9] C. Andrade, V. Menon, S. Ameen, and S. Kumar Praharaj, "Designing and conducting knowledge, attitude, and practice surveys in psychiatry: Practical guidance," *Indian J. Psychol. Med.*, vol. 42, no. 5, pp. 478–481, 2020.
- [10] CAST, *Universal Design for Learning Guidelines*, version 3.0. Wakefield, MA: CAST, 2024.
- [11] J. A. Silvestri and M. C. Hartman, "Inclusion and deaf and hard of hearing students: Finding asylum in the LRE," *Educ. Sci.*, vol. 12, no. 11, art. 773, 2022.
- [12] Assistive Technology Act of 1998, Pub. L. No. 105-394, 112 Stat. 3627 (1998).
- [13] A. M. Cook and J. M. Polgar, *Cook and Hussey's Assistive Technologies: Principles and Practice*, 3rd ed. St. Louis, MO: Elsevier, 2008.
- [14] L. Zhou, A. Parker, D. Smith, and N. Griffin-Shirley, "Assistive technology for students with visual impairments: Challenges and needs in teachers' preparation programs and practice," *J. Vis. Impair. Blind.*, vol. 105, no. 4, pp. 197–210, 2011.
- [15] P. Parette and M. Scherer, "Assistive technology use and stigma," *Educ. Train. Dev. Disabil.*, vol. 39, no. 3, pp. 217–226, 2004.
- [16] International Society for Technology in Education, *Will New Teachers Be Prepared to Teach in a Digital Age?* Santa Monica, CA: Milken Family Foundation, 1999.
- [17] D. Edyburn and J. Gardner, "Integrating technology into special education teacher preparation programs: Creating shared visions," *J. Spec. Educ. Technol.*, vol. 14, no. 2, pp. 3–17, 1999.
- [18] A. O. Onivehu, O. E. Ohawuiro, and B. J. Oyeniran, "Teachers' attitude and competence in the use of assistive technologies in special needs schools," *Acta Didact. Napocensia*, vol. 10, no. 4, pp. 21–32, 2017.
- [19] D. Chambers, "Assistive technology supporting inclusive education: Existing and emerging trends," in *Assistive Technology to Support Inclusive Education*, vol. 14, D. Chambers, Ed. Bingley, U.K.: Emerald, 2020, pp. 1–16.
- [20] Viberg, O., Cukurova, M., Feldman-Maggor, Y., Alexandron, G., Shirai, S., Kanemune, S., Wasson, B., Tømte, C., Spikol, D., Milrad, M., Coelho, R., & Kizilcec, R. F. (2025). What explains teachers' trust in AI in education across six countries. *International Journal of Artificial Intelligence in Education: Official Journal of the International AIED Society*, 35(3), 1288–131

Rethinking ‘Good’ in AI for Good: Communities as Judges of University-Led AI for Sustainable Development

Yang Bong
Birmingham Newman University
Birmingham, United Kingdom
y.bong@staff.newman.ac.uk

Neelima Sailaja
University of Nottingham
Nottingham, United Kingdom
Neelima.Sailaja@nottingham.ac.uk

Minjia Chen
University of Nottingham
Nottingham, United Kingdom
m.chen@nottingham.ac.uk

Devisakti Annamali
Universiti Sains Malaysia
Penang, Malaysia
devisakti@usm.my

Jin Tada
National Graduate Institute for
Policy Studies (GRIPS)
Tokyo, Japan
STARTandESSH@gmail.com

Leïla Cellérier
University of Geneva
Geneva, Switzerland
leila.cellerier@unige.ch

Muhammad Zahid Iqbal
Teesside University
Middlesbrough, United Kingdom
zahid.iqbal@tees.ac.uk

Vicente Pitogo
Caraga State University
Butuan, Philippines
vapitogo@carsu.edu.ph

Abstract

Artificial intelligence (AI) is increasingly promoted as an accelerant for the Sustainable Development Goals (SDGs), much of it advanced through the "AI for Good" discourse. Universities, as centres of AI research, deployers of applied systems, and institutions of public good are increasingly pivotal actors in this effort. Yet what constitutes “good” is often defined through researcher, funder, and institutional priorities rather than by the communities AI initiatives are intended to serve. Although the participatory turn in AI has expanded community involvement in system design, participation at the project level does not necessarily redistribute institutional authority over problem definition, data governance, and evaluation. Through an interdisciplinary conceptual synthesis drawing on the capability approach, participatory design, and development ethics, this paper reframes affected communities as receivers, doers, and crucially, judges of university-led AI for sustainable development. It argues that communities must possess meaningful authority to determine which capabilities matter and what constitutes successful development. The resulting institutional framework translates this proposition across universities’ research, teaching, and community-engagement missions and identifies implications for institutional governance, incentives, capacity building, and long-term partnerships. The framework contributes particularly to SDG 4, SDG 10, and SDG 17 by repositioning community judgement as a condition for legitimate university-led AI for Good.

Keywords

AI for Good, capability approach, participatory AI, evaluative authority, higher education institutions

1 Introduction

AI is now routinely presented as an accelerant for the Sustainable Development Goals [12, 16], and universities sit at the centre of that promise – producing the research, training the practitioners, and increasingly deploying the systems themselves. The claim such initiatives make is not merely technical but moral: that the work serves a good beyond the institution producing it. That claim invites a question the field has largely left to one side. Who defines ‘good’ in AI for Good? Universities are exceptionally well-placed to question this in their role of advancing AI for sustainable development [3, 17]. Rethinking ‘good’ presupposes a prior clarification of what “good” consists in – a clarification that more techno-solutionist paradigms that centre successful technology development over communities’ capabilities leave unspecified. The field’s dominant response to the “whose good” problem has been the ‘participatory turn’, widening community involvement in the design of AI systems [14]. Yet even accounts that centre communities’ participation [1] operate predominantly at the level of individual projects, leaving institutional structures that incentivise, fund, and evaluate AI projects underexamined [3]. This paper argues that participation, while necessary, is not sufficient if the ‘good’ in university-led AI projects is predefined by institutional actors before affected communities can meaningfully shape it. The distinction at issue is between voice and authority: participation gives communities a say in how a project proceeds [1, 2], while authority determines whether its purposes, its data and its criteria of success were theirs to set.

To address this gap, the paper develops an interdisciplinary framework for universities to specify how ‘good’ can be ascertained in university-led AI research. Through a conceptual synthesis drawing on scholarship in development ethics, participatory design, and the capability approach, the framework operationalises Sen’s capability approach [15] to reframe the ‘good’ of AI-enabled development, neither in the novelty of the systems deployed, nor in their declared social purpose, but in the expansion of the substantive freedoms communities have reason to value. Extending Bonvin and Laruffa’s [4] account of personhood from the individual to the community as a collective agent, the paper argues that affected communities must be positioned not only as receivers and doers of AI-enabled development but as its judges

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Information Society 2026, Ljubljana, Slovenia

© 2026 Copyright held by the owner/author(s).

<https://doi.org/10.70314/is.2026.ai4sdg.163>

— the source of valid claims about what counts as success. This reframing relocates evaluation from what a system produces to what it enables people to be and do.

2 The Participatory Turn in AI for Good

Since the late 2010s, “AI for Good” has consolidated from a loose aspiration into a recognised field, propelled by the UN’s Sustainable Development Goals, dedicated corporate initiatives, and a growing body of academic scholarship [6]. However, this rapid consolidation has also invited sustained critical inquiry. Recent scholarship has converged around three fundamental tensions, each questioning whose specific interests the field’s initiatives ultimately serve. The first concerns definition and power: critics observe that “good” is seldom specified and, when operationalised, tends to reflect the priorities of researchers, funders, and firms rather than the communities addressed [6, 10], producing a techno-solutionism in which deployment is mistaken for benefit [9]. The second is distributive: initiatives optimised for connectivity and data-rich environments are found to routinely fail those at the margins, so that access alone does not translate into benefit and existing inequalities are reinforced rather than relieved [11]. The third, and least examined, is epistemic. Lewis et al. [7] argue that mainstream AI rests on a scarcity mindset that treats intelligence as a discrete property of a self-contained agent and licenses extraction as the cost of progress. Mohamed et al. [9] reach a parallel conclusion from decolonial theory, calling for “reverse tutelage” to be made explicit so that knowledge flowing from communities to institutions is recognised as teaching rather than treated as an unacknowledged resource.

In response, a participatory turn has gained ground across the field, seeking to reposition affected communities from beneficiaries to co-designers [1, 14], increasingly through methods such as speculative co-design. Speculative co-design invites participants to imagine, sketch and debate technologies beyond existing constraints, treating the resulting artefacts as evidence of community values and possibilities that conventional consultation cannot reveal. Such approaches intentionally account for the community’s consideration of what is worth building, refusing or preserving [11]. Recent work in design justice - a movement that treats the design process itself, and not only what it produces, as a site of justice - holds that the practices, norms, and ownership structures of design must themselves be redistributed toward those bearing the burden of its harms [5]. Speculative co-design asks participants not what they want from a technology, but what they would build if the usual constraints were lifted. Recent work makes this lens operational for AI. Costanza-Chock [5] argues that overcoming the limits highlighted requires more than inclusive intent. Groups sketch, prototype and argue over imagined systems, and the artefacts they produce are read as evidence of what a community values rather than as product specifications. It surfaces what conventional consultation cannot, because presenting an existing tool only elicits reactions to someone else’s premise. Oluwatuyi et al.’s [11] work with the Ocean View community in Cape Town is illustrative where workshops produced speculative artefacts that centred linguistic and cultural sovereignty, surfacing what a community considers worth building, refusing, or preserving. Together, this body of work moves the field’s guiding question away from what AI can technically achieve toward whose knowledge counts, who participates, and who benefits, questions the following section sets out to operationalise.

Recent AI scholarship has begun to connect the capability approach with participatory design. London and Heidari [8] use the capability approach to evaluate whether AI systems expand an agent’s real freedoms, while identifying paternalism, coercion, deception, exploitation and domination as ways systems may undermine agency. Bondi et al.’s PACT framework [1] draws the participatory implication that “social good” cannot be specified in advance by researchers or funders, but requires affected communities to help identify the capabilities relevant to their flourishing. Bonvin and Laruffa’s [4] distinction between receivers, doers and judges (Figure 1) crystallises this final concern by positioning people not only as beneficiaries or contributors, but as agents entitled to determine what counts as worthwhile public action.

Taken together, these traditions provide complementary but incomplete accounts of AI for Good. The capability approach provides an evaluative standard centred on substantive freedoms, which participatory design translates into the field of technology and AI development. Yet without corresponding institutional levers, capability-oriented participation faces two risks: remaining confined to individual projects, or involving communities only after universities and funders have fixed the problem definition, data arrangements and success criteria, leaving evaluative authority institutionally unchanged. These perspectives suggest that the central issue is not a lack of ethical frameworks or greater participation, but a need for the institutional distribution of authority to define and evaluate the direction of development. This paper therefore extends the judge role from the individual to the community as a collective, though not necessarily unified, evaluative agent, and relocates it from project-level participation to the institutional structures through which authority is allocated. The framework identifies 1) problem definition, 2) data and model governance, and 3) evaluation as three levers across universities’ research, teaching and community-engagement missions, providing not another participatory method but an institutional account of when participation becomes evaluative authority.

COMMUNITIES’ ROLE	HOW UNIVERSITIES POSITION COMMUNITIES
 RECEIVER A vulnerable being who needs support	Recruited into studies designed elsewhere, for problems named elsewhere <i>over-emphasised</i>
 DOER An actor who contributes, and whose agency they have reason to value	Contributes data, local knowledge and testing to a system whose purpose is already fixed <i>counterfeit agency</i>
 JUDGE A political being with aspirations, who can define what counts as good	Consulted at best. Rarely holds authority over the questions, the terms, the measures, the values embedded and what counts as impact <i>seldom granted authority</i>

Figure 1: Receiver, Doer, Judge: how university research positions communities.

3 The University AI for Good Framework

Positioning communities as actual receivers, doers and judges [4] requires changes in universities at the institutional level. The

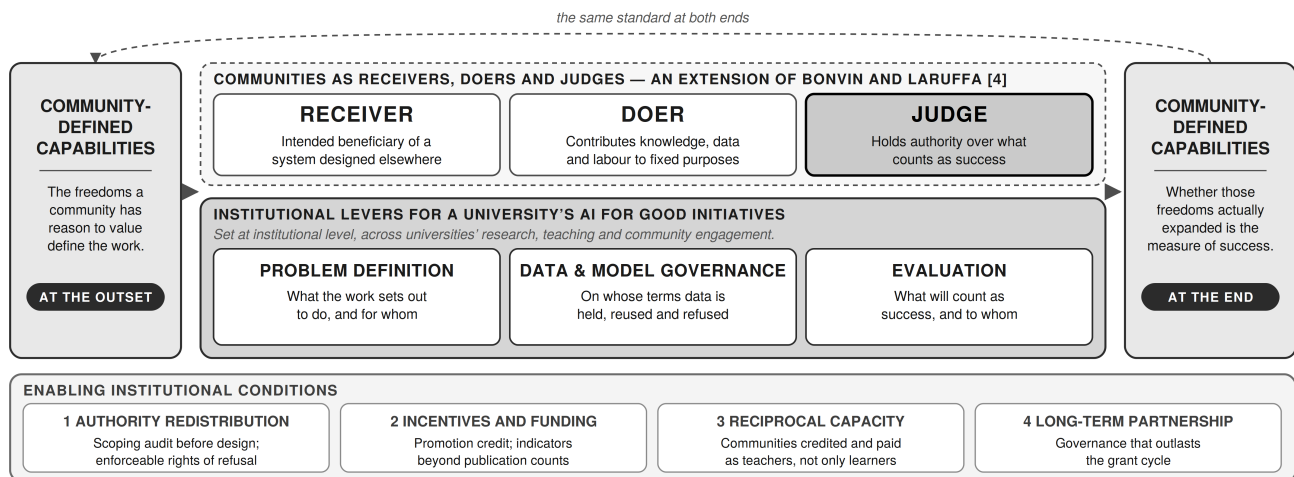


Figure 2: University AI for Good Framework: Communities occupy the receiver, doer and judge positions simultaneously. The judge position is constituted over three institutional levers set above the project, sustained by four enabling conditions.

three roles are institutional positions rather than dispositions, and each corresponds to a different kind of claim that a community can make. As receivers, communities are the intended beneficiaries of systems designed elsewhere, and the institutional question is one of access and equal distribution of benefits and harms. As doers, they contribute knowledge, data and labour to systems whose purposes are already settled, which is in the terrain of the participatory turn - co-design workshops, participatory data collection, value-sensitive design. As judges, they hold authority over what counts as success - a position routinely overlooked and not secured through participation as receivers or doers alone.

That authority, however, is not held at the project level; university AI project teams can influence it but cannot secure it through project-level participation alone. Whether a community can judge a university's AI work is settled by three levers (Figure 2). The first is problem definition: what an AI project sets out to do is shaped by funding calls and institutional research priorities before any community is approached. The second is data and model governance: the terms on which data are collected, and models are developed, held and reused follow from institutional agreements and legal defaults rather than being negotiated case by case. The third is evaluation: what will count as success is specified in advance by funders and institutional assessment regimes, in publications, patents and prototypes. A community can therefore be a full participant in an AI project whose success criteria were fixed before it began. This is why community participation cannot on its own deliver the judge role, and why the reorientation argued for here must be institutional.

Universities' missions span three main areas. In research, communities should co-define the problem and hold governance rights over the data a project generates, placing the first two levers within their reach. In teaching, they should be recognised as teachers and not only as learners, so that curricula are co-designed and community partners are credited and paid as co-instructors. In community engagement, partnership should be structured to outlast a single funding cycle, since evaluative authority exercised only within the life of a grant is authority over work that has already concluded. That such an arrangement is achievable is illustrated by existing practice. The Abundant Intelligences programme convenes Indigenous knowledge-holders

and community organisations as co-investigators rather than informants across a six-year, multi-institution consortium, locating authority over what is built - and over the knowledge that justifies building it - with communities from the outset [7].

4 From Framework to Institutional Practice

The paper therefore argues for a capability-centred, community-led reorientation of AI for Good - one whose implementation requires moving beyond siloed adoption to strategic change across institutional levels. The analysis identifies four sites at the institutional level where the judge role is won or lost: the governance of problem definition and AI governance, the incentives that reward academic work, the direction capacity is built in, and the time horizon of partnerships.

4.1 Authority over problem definition and AI governance

Operationally, community-defined good means moving communities beyond the receiver role alone and into those of doer and judge. The shift that matters is not from consultation to co-design, since co-design remains the terrain of the doer; it is from contributing to a project to holding authority over its terms, which follows from recognising affected populations as the primary domain experts on their own conditions [5]. Two mechanisms carry that authority. The first is a scoping audit, conducted before a project is designed, establishing who defines the problem, owns the data and governs the system, and testing the proposal against the five failure modes London and Heidari [8] identify - paternalism, coercion, deception, exploitation and domination - each a way a system can serve a declared good while withholding agency from those it serves. The second is refusal: governance agreements should state not only what rights communities hold over their data, but what refusal triggers and when it can be exercised, including declining at scoping, redirecting the work, withdrawing data access, or ending the partnership, with an escalation route outside the research team and terms settled at the outset for what the community keeps if the partnership ends. Authority that cannot be exercised against the institution granting it is not authority but consultation on a longer timeline.

4.2 Incentives for community-engaged work

Participatory, capability-centred work is more time and resource-intensive than deployment-first research [1], and prevailing evaluation criteria – publications, grants and speed to deployment – reward the latter. Unless the surrounding incentives change, a single team's commitment to community-engaged work is re-absorbed by the pressures around it. This is how the participatory turn stalls at project level: teams can redistribute authority within their own projects, but not the criteria against which those projects are judged, which promotion committees, assessment regimes and funding cycles set above them. Three changes follow: standing for community-engaged work in promotion criteria; funder support for multi-year partnerships in which communities hold governance rights from the outset; and indicators reaching beyond publications, patents and prototypes to community capability, partnership quality and public value.

4.3 Capacity built in both directions

Adopting a capability lens allows universities to enquire beyond what immediate local problems AI research could resolve or be developed for a community, to seek a more grounded understanding of the capabilities that a said community values and how technology might expand them [1, 15]. Two consequences follow. Communities should be approached not through a deficit lens but as holders of local knowledge and problem-solving capacity on which AI builds rather than substitutes; and if knowledge is to flow in both directions, the reverse tutelage scholars describe requires that communities be recognised and credited as teachers, not only as learners [9]. In practice, this means embedding capability-based AI literacy into curricula co-designed with affected communities rather than standalone access initiatives, and bringing community partners into teaching as co-instructors who are recognised, paid and credited [3] while training students in the co-design methods.

4.4 Long-term partnerships and cross-institutional infrastructure

Projects that collapse when a grant ends damage public trust and deepen the very disillusionment that participatory work seeks to overcome. Sustainability is therefore a condition of doing this work ethically rather than an operational afterthought, and it bears on the evaluative lever directly: a cycle that closes before a community can assess what was built leaves judgement with the funder by default. Universities can use their convening power to establish funding outlasting the grant cycle, pooling institutional, municipal, philanthropic and private mechanisms to maintain infrastructure and fairly compensate the knowledge-holders whose contributions make these systems work. Shared platforms for exchanging methods, governance frameworks and community-centred innovations let arrangements developed in one setting be adapted in another [13] – provided the resulting commons documents local practice on terms its holders set [9].

5 Conclusion

The paper proposes an institutional framework through which universities can position communities not only as beneficiaries and contributors, but as judges whose values and capabilities inform problem definition, data and model governance, and evaluation. The contribution of the paper is twofold: it extends Bonvin and Laruffa's [4] account of persons as receivers, doers and judges from the individual to the community as a collective agent, and

it locates the judge role at institutional rather than project level, where the levers that constitute it sit. The framework speaks to SDG 4 through curricula co-designed with the communities they serve, SDG 10 by treating the distribution of evaluative authority as itself a form of inequality, and SDG 17 by making partnership a governance arrangement.

Rethinking 'good' at a time when AI is claimed to serve a broad range societal, environmental and economical good remains a timely and important one. Within universities, every AI project encodes an answer in who defined the problem, whose data trained the system, and who was in the room to evaluate success. Without deliberate choices, prevailing incentives supply the answer: good becomes whatever is fundable, publishable and measurable. Inaction is hence not neutral. Beyond the moral argument, universities also have institutional interest at stake: their public funding, access to communities, and research trust depend on it. If universities cannot say who benefits from their AI work, they will struggle to sustain their claim that such work serves the public good.

References

- [1] Elizabeth Bondi, Lily Xu, Diana Acosta-Navas, and Jackson A Killian. 2021. Envisioning communities: a participatory approach towards ai for social good. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 425–436.
- [2] Yang Bong, Lewis Cameron, Gary Loh Chee Wyai, Tariq Zaman, and John Harvey. 2025. 'can my voice be heard?': taking non-weird voices seriously in digital marketplaces. In *Proceedings of the ACM SIGCAS/SIGCHI conference on computing and sustainable societies*, 297–310.
- [3] Yang Bong, Sharina Yunus, Rashikala Weerawarna, Roman Senkerik, Rui Tan, Tehreem Masood, Thomas Leoni, Thomas Wawn, and Vanessa Camilleri. 2025. Universities' role in the artificial intelligence innovation system for sustainable development. *Asia Europe Foundation Publications*.
- [4] Jean-Michel Bonvin and Francesco Laruffa. 2018. Human beings as receivers, doers and judges. the anthropological foundations of sustainable public action in the capability approach. *Community, Work & Family*, 21, 5, 502–518.
- [5] Sasha Costanza-Chock. 2020. *Design Justice: Community-led Practices to Build the Worlds We Need*. The MIT Press.
- [6] Josh Cows. 2021. 'ai for social good': whose good and who's good? introduction to the special issue on artificial intelligence for social good. *Philosophy & Technology*, 34, Suppl 1, 1–5.
- [7] Jason Edward Lewis, Hēmi Whaanga, and Ceyda Yolgörmez. 2025. Abundant intelligences: placing ai within indigenous knowledge frameworks. *AI & Society*, 40, 4, 2141–2157.
- [8] Alex John London and Hoda Heidari. 2024. Beneficent intelligence: a capability approach to modeling benefit, assistance, and associated moral failures through ai systems: aj london and h. heidari. *Minds and Machines*, 34, 4, 41.
- [9] Shakir Mohamed, Marie-Therese Png, and William Isaac. 2020. Decolonial ai: decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33, 4, 659–684.
- [10] Jared Moore. 2019. Ai for not bad. *Frontiers in Big Data*, 2, 32.
- [11] Rachel Oluwatuyi, Vimlan Pillay, Jasmine Mazwi, Alvin Castro, Jaydon Farao, Shreyas Srivatsa, Naveen Bagalkot, and Melissa Densmore. 2026. Collectively reimagining artificial intelligence with marginalized communities. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 1–21.
- [12] Iván Palomares et al. 2021. A panoramic view and swot analysis of artificial intelligence for achieving the sustainable development goals by 2030: progress and prospects. *Applied Intelligence*, 51, 9, 6497–6527.
- [13] Neelima Sailaja, Joel E Fischer, Rishub Jain, Neha Kumar, Simone Stumpf, Min Kyung Lee, Heloisa Candello, and Raquel Iniesta. 2026. Rai@ chi: responsible and human centred ai across borders. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, 1–3.
- [14] Devansh Saxena et al. 2025. Emerging practices in participatory ai design in public sector innovation. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–7.
- [15] Amartya Sen. 1999. *Development as Freedom*. Vol. 6. Oxford university press Oxford.
- [16] Ricardo Vinuesa et al. 2020. The role of artificial intelligence in achieving the sustainable development goals. *Nature communications*, 11, 1, 233.
- [17] Timothy Wawn, Tehreem Masood, Oana M Driha, Yenni Astuti, and Yang Bong. 2024. Positioning for success: how universities in developing countries can thrive in the ai era by focusing on sustainable development. *Asia Europe Foundation Publications*.

Artificial Intelligence Readiness in Brazilian Schools: A Nationwide Survey of Teachers and Educational Challenges

Bruno Florentino
Department of Computer
Science
University of São Paulo
São Carlos, São Paulo, Brazil
brunorf1204@usp.br

André de Carvalho
Department of Computer
Science
University of São Paulo
São Carlos, São Paulo, Brazil
andre@icmc.usp.br

Camila Sestito
Department of Computer Science
Federal University of
Technology-Paraná (UTFPR)
Cornélio Procópio, Paraná, Brazil
camilasestito@utfpr.edu.br

Wellington Cruz
Educational President
Instituto Significare
São Paulo, São Paulo, Brazil
prof.well@significare.org.br

Robson Bonidia
Department of Computer Science
Federal University of
Technology-Paraná (UTFPR)
Cornélio Procópio, Paraná, Brazil
bonidia@utfpr.edu.br

Abstract

Artificial Intelligence (AI), particularly Generative AI, has emerged as a transformative force in education, with the potential to personalize learning, support inclusion, and improve teacher productivity. Brazil has begun to address this through initiatives like the Brazilian AI Plan (PBIA 2024-2028) and the National Digital Competence Framework for Teachers, yet no official curriculum defines how AI should be integrated into basic education, leaving a gap between policy and classroom reality. This study surveyed 346 Brazilian basic education teachers nationwide to assess AI readiness across three dimensions: Teaching and Learning with Digital Technologies, Digital Citizenship and Professional Development, contextualized with data from the 2024 Brazilian School Census, IBGE, OECD, and international experiences (Canada, Hong Kong). The results show strong enthusiasm: 97.3% of teachers recognize at least one educational benefit, especially interactive material creation (80.6%), lesson planning (80.2%), personalized assessment (68.6%), and inclusive practices (65.5%). More than 80% want continuous AI training, preferring practical, collaborative online courses. However, the main barriers remain: inadequate training (43.4%), poor

infrastructure and internet access (43.0%), lack of technical support (41.9%), and limited time (34.5%), confirmed by national disparities in school connectivity and equipment. Rather than resistance, the findings reveal a disconnect between AI policy and school conditions, mapping the structural and governance challenges for the equitable adoption of AI in Brazilian schools.

Keywords

Artificial Intelligence, K-12 Education, Brazilian Schools

1 Introduction

New technologies tend to prioritize the Global North in their development and application, a pattern that persists with Generative Artificial Intelligence (GAI) [1]. In education, GAI offers notable benefits along with risks that must be carefully weighed [8]. For teachers, it supports the creation of teaching materials and instructional design [2]–[5], [8], and assists with administrative tasks such as the drafting of proposals and recommendation letters [8]. It also enables automated assignment correction with instant feedback [2]–[4], [8], generates visual content to clarify complex concepts [2], [4], and produces personalized texts and activities based on student performance [2], [4], [5]. However, the effectiveness of these applications remains debated, as GAI models show inconsistent accuracy and quality across tasks [2], [5].

For instance, GPT-4 achieves only 77% accuracy in programming tasks [2], and performs unsatisfactorily in advanced mathematics [5]. Models also embed biases

¹Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Information Society 2026, 5–9 October 2026, Ljubljana, Slovenia
© 2025 Copyright held by the owner/author(s).
<https://doi.org/10.70314/is.2026>

reflecting their training data [2], [5], [8], compounded by low interpretability [2], [5]. Because GAI output synthesizes training data rather than producing genuine innovation, a “pseudo-imagination” [2], its classroom use risks generic results that discourage creativity and authenticity [2], while complicating efforts to distinguish student- from AI-generated work and raising academic integrity concerns [2], [8].

Beyond academics, GAI lacks the emotional communication that human teachers can provide [3]. A growing preference of students for AI interaction risks weakening student-teacher relationships and educators' ability to identify emotional and pedagogical needs [3]. Taking this into account, our study examines AI education in Brazil, a country facing significant barriers to technology access, ranking 51st among 127 economies in a global connectivity index [6]. These barriers stem from structural inequalities; e.g., [7] found a strong correlation between internet speed and wealth, with the greatest disparities in Rio de Janeiro, where connectivity in favela regions lags far behind the rest of the city. Such inequalities predate AI but tend to intensify as it integrates into society.

Given Brazil's technological inequalities and the potential impact of GAI's on K-12 education, understanding the perceptions, concerns, and expected benefits of Brazilian educators is essential information that also informs the broader Latin American context, given Brazil's position in regional development [9]. Therefore, we propose a quantitative study investigating the opinions of K-12 teachers about AI and GAI in Brazil, exploring their expectations, perceived impacts, and implementation challenges. To our knowledge, this is the first nationwide survey that identified classroom teachers' actual needs regarding AI use.

2 Methodology

The Brazilian Ministry of Education published the **Framework of Teachers' Digital Competencies** [10], structured into three dimensions and ten competencies to support teachers' continuous professional development on digital technologies in K-12 education. “**Teaching and Learning with the Use of Digital Technologies**” covers pedagogical practice, curation and creation, data analysis, and inclusive practice areas well aligned with AI-based technologies. “**Digital Citizenship**” covers responsible, safe, and critical use, addressing data privacy and technology's impact on student well-being. “**Professional Development**” covers continuous professional development, communication and collaboration, and use of digital resources for management.

We adopted this framework as the conceptual basis for assessing teachers' perceptions of the topic. Accordingly, a 35-item quantitative questionnaire was developed to systematically investigate the thematic dimensions defined by the framework. These elements were also derived from an initial questionnaire based on a literature

review, which was initially administered to four educators from different regions of the country, leading to the final version of the material, which can be accessed in the Final Form – Artificial Intelligence in the Classroom²

Data were collected from October 28, 2024 to 2025, using an online form (Google Forms), with the aim of a wide reach across different regions of Brazil. The sample consisted of 346 teachers from various regions. Considering the total population of approximately 2.4 million active teachers in Brazilian basic education, according to the 2024 School Census [10], this provides a confidence level of 90% with a margin of error of 5%.

2.1 Sample

Understanding the characteristics of the sample is essential to interpret the data in context and to recognize the potential limitations of the study. Regarding the age range of the participants, the most representative group is between 45 and 54 years old, which corresponds to 33.3% of the sample. This is followed by the age group 35 to 44 (27.7%) and the age group 55 to 64 (21.4%). In parallel, the 2024 School Census [10] indicates that approximately 35% of Brazilian teachers are between 40 and 49 years old. Another relevant aspect concerns the length of experience in the field of education. The majority of the participants (44.2%) have more than 20 years of experience, followed by 28.1% with 11 to 20 years of experience. Additionally, 12.4% reported having 6 to 10 years of experience, while 10.1% are in the range of 1 to 5 years. The least represented groups are those with less than 1 year of experience (1.7%) and those who do not directly work in education (3.5%).

The sample also provides information on the level of education at which professionals work. Approximately 29.5% of the respondents are associated with high school, 27.1% work in Elementary School I (grades 1–5), and 16.7% in Elementary School II (grades 6–9). These data help to understand the diversity of educational contexts represented in the study. We also analyze the affiliation of the institutions with the government. Public schools in the state represent the majority of the survey responses, accounting for 37.6% of the total, followed by public schools in the municipal district at 28.7%. In third place are private institution educators, comprising 17.1% of respondents. The participation of professionals affiliated with federal public schools is 9.3%, while those working in universities represent 3.5%. Furthermore, 3.8% of the participants reported working in other contexts.

3 Results and Discussions

The findings show that 97.3% of educators are interested in the potential of AI for teaching and learning, although effective implementation faces challenges such as teacher training, infrastructure, and technical support. Most educators report only basic AI knowledge (80.3%),

² <https://forms.gle/aTRJ4LE257EugDR56>

limiting their use beyond simple tasks. Still, more than 80% want continuous AI training, preferring online courses with videos, quizzes, and collaborative spaces that indicate openness to learning but a need for structured, practical materials, since generalist courses often lack applicability to each teacher's context and subject area.

We observed that most of the respondents have limited knowledge on the subject: approximately 69.4% reported having a basic understanding, 9.7% stated that they had only heard about the topic, and 1.2% indicated not knowing at all. In contrast, 17.8% reported having advanced proficiency, and 1.9% considered themselves experts. Next, aligned with the competencies of the first dimension, we analyze the participants' perceptions about the potential benefits of AI in teaching. In this question, respondents who recognized the potential of AI could select multiple areas in which it could contribute, while those who did not identify benefits selected only that single option. Taking this into account, 97.3% of the participants indicated at least one positive aspect among those listed.

The options cited most frequently were the creation of interactive content (80.6%) and lesson planning (80.2%). Other benefits mentioned included personalized assessment (68.6%), the possibility of inclusion (65.5%), and the identification of student learning difficulties (47.7%). The study also revealed that teachers are interested in using AI for the analysis of educational data, perceiving it as a valuable resource for pedagogical practices.

In this regard, 79.8% of the participants highlighted the identification of learning trends as the main benefit. In the field of inclusive practices, 71.3% pointed to the promotion of personalized learning strategies, while 70.9% recognized the importance of integrating assistive technology resources. Other benefits mentioned include the creation of collaborative learning environments (60.5%) and the monitoring of student progress, with a focus on identifying areas that require special attention (53.5%). Only 0.8% stated that they do not believe in the potential of AI to promote inclusion.

Moving on to the second dimension, digital citizenship, the results indicate that 78.3% of educators consider it essential to promote students' education on ethical issues related to AI, recognizing this aspect as crucial for the critical use of these tools. Furthermore, 73.6% emphasized the importance of teaching students how to critically evaluate AI-generated information.

Other aspects were also widely highlighted by the participants. Among them, 70.2% emphasized the importance of reflecting on technological dependence and its impacts in the classroom. Furthermore, 64.3% considered it essential to ensure that teachers receive adequate training in ethics and safety in the use of AI. Finally, 61.2% stressed the need to ensure transparency in decisions made by AI systems, as well as in the identification and mitigation of their biases.

Regarding the responsible use of AI by students, both inside and outside the classroom, 39.1% of teachers reported that students demonstrated some knowledge of the topic but did not use these tools frequently. Meanwhile, 22.5% indicated that students actively use AI in their daily academic activities. In contrast, 13.6% reported that most students do not use AI in their practices.

The third dimension refers to professional development. In this area, 79.8% of teachers highlighted the importance of facilitating access to online platforms and learning materials about AI. Furthermore, 72.5% indicated that the provision of courses and content tailored to individual levels and knowledge needs is relevant. Other important aspects include the suggestion of courses and training aligned with the professional trajectory (65.5%), the promotion of collaboration and experience exchange between educators (65.1%), and the possibility of using AI for performance analyses to identify areas for development (57.4%). Only 0.4% of the respondents do not believe that AI can contribute to continuous professional development.

However, the majority of respondents (50.8%) stated that they have not participated in any AI courses in education but are interested. Another 25.2% reported having participated in more than one training session, while 23.6% attended a single course and 0.4% expressed no interest. Teachers also recognize that AI can facilitate communication and collaboration, as well as support professional development. The main highlight was the use of AI to locate and share teaching materials available online, indicated by 72.1% of the respondents. Other benefits mentioned include facilitating the creation and participation in educator groups (68.6%) and supporting the co-creation of digital resources and educational content.

Regarding the level of support and encouragement for the use of AI, the majority (31.8%) believe that the school offers some encouragement, but it is still insufficient. Another 26.4% indicate that the school leadership and the school community are aware but do not promote its use, while 26.0% report that there is no significant technology-oriented culture. Only 13.0% reported strong leadership support, and 2.8% were unable to respond.

4 Discussion

Regarding inclusion, the study indicated that respondents recognize the potential of AI as an ally in creating teaching materials that are more accessible and compatible with the diverse needs of students, with the content personalization proposal widely acknowledged in other studies [2], [8]. According to the 2024 School Census [10], special education enrollments exceeded 2.1 million, representing an increase of 58.7% in the past five years. This growth reinforces the need highlighted by study participants for the creation of interactive content (80.6%), personalized activities (84.1%), and the

development of accessible resources (81%), demonstrating the importance of inclusive strategies 'supported by technologies such as AI.

To contextualize this scenario, we can analyze student-to-class ratios from the 2024 School Census [10] and teacher working hours, according to the OECD's Education at a Glance 2024 report [11]. National data indicate that the average number of students per class is 15.4 in early childhood education, 20.9 in the lower grades of primary education, 25.9 in the upper grades, and 28.9 in secondary education. These numbers show that class sizes tend to increase with the education level, requiring teachers to manage greater diversity, and, consequently, leaving less time to address individual student needs.

Beyond AI's benefits, it is essential to understand the practical obstacles to its implementation. The lack of technical support, reported by 41.9% of the participants, is a significant barrier, reflecting the shortage of school personnel able to assist with computers, projectors, digital whiteboards, and internet connectivity. These challenges extend beyond training to structural problems that limit technology adoption. The 2024 School Census [10] indicates that 85.9% of basic education schools have broadband internet access.

However, the Connected Education program [12], which evaluates connection quality, reveals regional disparities: In São Paulo, part of the Southeast, the region with the best indicators, only 4,964 of 19,006 schools have adequate-speed connections with satisfactory measurement devices, a 26.1% connectivity rate. Other infrastructure data from the 2024 School Census [10] reinforce these limitations: only 70% of schools have multimedia projectors and around 21% have digital whiteboards, restricting real-time projection of content, including AI-generated materials. Approximately 56% of schools have desktop computers, with even lower availability of laptops and tablets, preventing practical, teacher-supervised activities exploring AI tools.

Finally, this study relies on self-selected, self-reported survey data; the reported 90% confidence level should be interpreted with caution given the non-probability sampling design. Additionally, since most respondents report only basic AI knowledge, their perceptions may not fully reflect AI's actual classroom impact, and no independent measure of AI competence or learning outcomes was collected.

5 Conclusions

In our view, the expression "AI for all" means that every individual, regardless of geographical location, financial resources, or prior knowledge, should have access to AI and its benefits. Taking this into account, we sought to understand how this principle translates to the reality of Brazilian K–12 education and, most importantly, to listen to those who will directly face the challenges and opportunities in classrooms: teachers themselves. Our

findings reveal that enthusiasm for AI is not enough on its own: turning "AI for all" into reality requires coordinated action, continuous teacher training, infrastructure investment, and public policies that address the structural and socioeconomic disparities that still shape Brazilian schools.

6 Partnership and Data Availability

This survey was conducted by UTFPR-CP, InteliGente Project, and Instituto Significare. The complete report is available at <https://doi.org/10.48550/arXiv.2512.23834>.

Acknowledgements

This work was funded by The Spencer Foundation (Grant Agreement 202600243) and the Canadian International Development Research Centre (IDRC, Grant Agreement 109981), with support from the UK government's Foreign, Commonwealth and Development Office. Bruno Florentino was funded by the FAPESP, grant #2024/00830-8. Dr. Bonidia received funding from the Regional Fund for Digital Innovation in Latin America and the Caribbean (FRIDA-LACNIC).

References

- [1] D. Yu, H. Rosenfeld, and A. Gupta, "The AI divide between the Global North and the Global South," World Economic Forum, 2023. [Online].
- [2] U. Mittal, S. Sai, V. Chamola, and D. Sangwan, "A comprehensive review on generative AI for education," IEEE Access, vol. 12, pp. 142733–142759, 2024, doi: 10.1109/ACCESS.2024.3468368.
- [3] X. Qiu and S. Zhang, "Application analysis of generative artificial intelligence in basic education," in Proc. 2024 7th Int. Conf. E-Business, Inf. Manage. Comput. Sci. (EBIMCS), 2025, pp. 41–46, doi: 10.1145/3723420.3723428.
- [4] J. Zhang and D. Sun, "A systematic review of generative artificial intelligence in education," in Proc. 7th Int. Conf. Comput. Sci. Technol. Educ. (CSTE), 2025, pp. 552–556, doi: 10.1109/CSTE64638.2025.11092288.
- [5] Q. Lang, M. Wang, M. Yin, S. Liang, and W. Song, "Transforming education with generative AI: Key insights and future prospects," IEEE Trans. Learn. Technol., vol. 18, pp. 230–242, 2025, doi: 10.1109/TLT.2025.3537618.
- [6] Portulans Institute, "Network Readiness Index 2025: Brazil" 2025. [Online]. Available: <https://networkreadinessindex.org/country/brazil/>.
- [7] N. Gozzi, N. Comini, and N. Perra, "Bridging the digital divide in six Brazilian cities," EPJ Data Science, vol. 13, no. 69, 2024, doi: 10.1140/epjds/s13688-024-00508-8.
- [8] D. T. K. Ng, E. K. C. Chan, and C. K. Lo, "Integrating generative AI in education," Computers and Education: Artificial Intelligence, vol. 8, Art. no. 100373, 2025, doi: 10.1016/j.caeai.2025.100373.
- [9] A. Tsang and N. Fu, "Brazil's economic development," HKTDC Research, 2025. [Online]. Available: <https://research.hktdc.com/en/article/MTk5MTYxNDg1Mg>
- [10] BRASIL. INEP, "Censo Escolar 2024: Resumo Técnico," Brasília, Brazil, 2025. [Online].
- [11] OECD, Education at a Glance 2024. Paris, France: OECD Publishing, 2024, doi: 10.1787/c00cad36-en.
- [12] BRASIL. Ministério da Educação, "Programa de Inovação Educação Conectada," Brasília, Brazil, 2025. [Online]. Available: <https://medidor.educacaoconectada.mec.gov.br>

Beyond AI Governance: Digital Space Resilience and the Bad Actor Problem

Sebastian Hallensleben

CPDiF – Centre for a Positive Digital Future
Aachen, Germany
sebastian.hallensleben@cpdif.org

Francesca Bosco

CPDiF – Centre for a Positive Digital Future
Geneva, Switzerland
francesca.bosco@cpdif.org

Abstract

AI governance has become the default response to AI-induced risks. However, it is limited in that it assumes actors can be reached by rules and standards, and therefore offers little against bad actors who ignore rules and standards. A substantial share of AI-induced risks is better and more sustainably addressed by upgrading the protocols and infrastructures of the digital space that form the context for both benign and malicious use of AI. Mechanisms for identity, provenance, attribution, trust tracking and accountability determine what a bad actor with a capable model can actually do. To inform such an approach, we draw lessons from the field of cybersecurity that has spent three decades on this problem: (1) defence must assume compromise rather than exclusion; (2) attacker economics matter more than attacker intent; (3) liability and incentive allocation shape outcomes; and (4) protocol-level measures are often more effective than user-level policy. We map these lessons onto AI and derive a research agenda for a resilient digital space in the context of the recent Montreal Memorandum, which sets out an overall path for an “upgrade” of the digital space. We also survey the spectrum of disciplines and communities that have started to converge within the Digital Trust Convention and propose a pathway for incubating and catalysing such collaboration in support of the Sustainable Development Goals, in particular SDGs 9, 16 and 17. Within this broader approach, AI governance still matters but becomes a narrower and considerably more tractable task.

Keywords

AI governance, digital trust, resilience, cybersecurity economics, digital identity, content provenance, accountability, Sustainable Development Goals

1 Introduction

Artificial Intelligence (AI) governance has become the default institutional response to AI-induced risks. Prominent examples include the European Union’s AI Act that establishes

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Information Society 2026, 5–9 October 2026, Ljubljana, Slovenia

© 2026 Copyright held by the owner/author(s).

<https://doi.org/10.70314/is.2026.ai4sdg.1>

a comprehensive, risk-based regulatory regime [1]; the U.S. National Institute of Standards and Technology that offers a voluntary risk management framework [2]; both part of a dense landscape of principles, codes of conduct and technical standards. These instruments have diversified in recent years and differ in scope and enforceability, but they still share a fundamental assumption: that the actors whose behaviour is to be steered can be reached by rules and standards, and, where necessary, sanctioned. For many AI-induced risks, this assumption does not hold. Fraud, disinformation operations, social engineering, non-consensual synthetic imagery and the manipulation of markets and elections are perpetrated by bad actors who, by definition, ignore rules. Misuse of AI is not hypothetical: forecasting work anticipated malicious use as early as 2018 [3], and joint analysis by Trend Micro, UNICRI and Europol documented operational criminal use – including password guessing, CAPTCHA breaking and voice cloning – by 2020 [4]. Nor do technical compliance measures embedded in models deter such actors: the safety alignment of large language models can be removed by fine-tuning at a cost of well under one dollar [5], and the safety training of openly released model weights can be undone reliably [6]. Since capable open-weight models proliferate globally, AI governance leaves the bad actor problem mostly unaddressed.

We argue that a substantial share of AI-induced risks is better and more sustainably tackled by upgrading the protocols and infrastructures of the digital space that form the context for both benign and malicious use of AI. Mechanisms for identity, provenance, attribution, trust tracking and accountability determine what a bad actor can do: how many convincing synthetic personas can be operated, how far deceptive content can spread, and what it costs when an operation is exposed. This argument connects to the Sustainable Development Goals (SDGs) in several ways: A digital space in which actors cannot be held accountable, in which reliable information is drowned out by synthetic content, and in which participation can be simulated at scale undermines SDG 16 (Targets 16.6, 16.10). The mechanisms we describe are infrastructure in the sense of SDG 9 (Target 9.1). The breadth of disciplines and sectors required maps onto the multi-stakeholder partnerships of SDG 17 (Targets 17.16, 17.17). The Global Digital Compact adopted by the UN likewise commits governments to fostering an inclusive, open, safe and secure digital space and to strengthening information integrity [7]. – The remainder of this paper distills four lessons from three decades of cybersecurity research and practice (Section 2), maps them onto a research agenda for digital space

resilience (Section 3), situates that agenda within the Montreal Memorandum and proposes a pathway for incubating collaboration (Section 4). Section 5 concludes with the narrower, more tractable role that remains for AI governance.

2 Four Lessons from Cybersecurity

Cybersecurity has confronted the situation described above almost since the emergence of the Internet: adversaries who cannot be reached by rules, capabilities that cannot be contained, and defenders who must nevertheless keep systems open and usable. Its accumulated experience – successes as well as failures – is the closest available body of evidence for designing a resilient digital space in the age of AI. We highlight four lessons and map each onto AI.

2.1 Defence Must Assume Compromise

For its first two decades, network security was organized around the perimeter: a hardened boundary meant to keep attackers out, with trust granted to everything inside. The model failed repeatedly, because a breach yielded unhindered lateral movement. The response, now codified as *zero trust architecture*, inverts the assumption: no trust is granted based on network location or ownership, every access request is verified, and the architecture is designed to limit the blast radius of a compromise [8]. - The AI analogue of perimeter defence is capability containment: keeping powerful models out of malicious hands through licensing, export controls, usage policies and refusal training. Each has value, but the evidence above – cheap removal of safeguards [5], [6] and global proliferation of open weights – indicates that containment can delay, not prevent. The zero trust lesson is to design under the assumption that bad actors already possess capable models. The question is not whether a bot, deepfake or influence operation can be created but what the environment allows it to achieve: how easily bots acquire standing, how far content propagates, and how quickly harms are detected and contained.

2.2 Attacker Economics Matter

Security economics, established as a field around the turn of the millennium, showed that security failure is often caused by misaligned incentives, and that these are as important as technical measures in achieving secure systems [9]. Defenders rarely change attacker intent; they change the attacker's business case. Classes of attack which thrive are those that are cheap and profitable: the CyberPeace Institute's analysis of cyberattacks against healthcare during the COVID-19 pandemic attributed their persistence to their being "low-risk, high-reward" crimes for the perpetrators [10]. - Generative AI has transformed the calculus on the attacker's side: the cost of producing content, personas and interactions has shrunk by orders of magnitude. Restoring a defensible asymmetry therefore requires interventions that e.g. raise the cost of acquiring an identity that passes for a person, raise the cost of keeping accounts alive, or raise the expected loss when

attribution succeeds. This is effective against criminal, commercial and geostrategic operators alike. - Furthermore, resilience depends on the economics faced by defenders. Sustainable interventions reduce the long-term costs of verification, recovery, coordination and institutional trust maintenance. Evaluating resilience mechanisms should consider the distribution of costs across the entire ecosystem.

2.3 Liability and Incentive Allocation

A second insight of security economics concerns the allocation of losses. Systems tend to fail when the actors best positioned to prevent harm are not also bearing its consequences; misaligned incentives and externalities contribute to persistent insecurity [9]. Protective investment occurs when responsibility and incentives are aligned with the capability to mitigate risk. In digital ecosystems, however, mitigation capability is rarely concentrated in a single actor but distributed across layers of infrastructure, services and applications. - The harms of AI misuse represent systemic externalities: they fall on defrauded individuals, manipulated communities and destabilized institutions, while mitigation levers—including interfaces, defaults, identity, provenance and amplification—remain under the control of operators whose commercial incentives may not align with public-interest. These mitigation choices are not merely technical decisions but governance choices shaped by standards, procurement practices, platform policies, obligations and oversight mechanisms. - Accountability and liability mechanisms may serve as proxies for encouraging investment. However, such mechanisms must be designed to avoid disproportionate burdens on open research, smaller actors and civil society.

2.4 Protocol-Level Beats User-Level

Decades of user education did not secure the web. Herley showed that users' rejection of security advice is economically rational: the aggregate cost in time and effort of following advice such as scrutinizing URLs exceeds, by orders of magnitude, the losses the advice would prevent [11]. What actually shifted the security of everyday web use was infrastructure: transport encryption became the default because browsers changed their interfaces and requirements, certificates became free and automated, and platforms demanded HTTPS – within roughly one year, the number of top websites serving HTTPS by default nearly doubled [12]. Users did not become drastically more vigilant; the ecosystem stopped depending on their vigilance. - The prevailing response to AI-generated deception repeats the pre-protocol pattern: media-literacy campaigns and appeals to "check the source" place the burden on the finite human attention that the attack exploits. The lesson is not that literacy is worthless, but that it cannot be relied on at a societal scale. Authenticity signals, provenance information and identity assurance must instead become ambient properties of the digital environment – present by default, verified by infrastructure, and shown to users only as simple, trustworthy cues, in the way the browser padlock abstracted away the cryptography beneath it.

3 A Research Agenda for Digital Resilience

Applying these lessons, we identify five mechanism families whose design determines what a bad actor can actually do. We regard them as a research agenda and as elements of an emerging public interest digital infrastructure.

3.1 Identity and Personhood

The digital space must restore the asymmetry that a human being is scarce while a synthetic persona is not. Privacy-preserving credentials with which a holder proves that they are a real person without revealing which one have been proposed because legacy countermeasures such as CAPTCHAs are no longer adequate against capable AI [13]. In Europe, the European Digital Identity framework creates a regulatory and technical basis on which such schemes could operate [14]. The Montreal Memorandum generalizes the design goal as *accountable, authentic pseudonymity*: identities that protect privacy and legitimate role separation while limiting large-scale automated duplication of participation [15]. Open questions include equitable global issuance, so that resilience does not become exclusion – a direct concern of SDG Target 16.9 – as well as the economics of Sybil resistance, revocation and recovery, and interoperability across jurisdictions.

3.2 Provenance and Authenticity

Content provenance standards, most prominently the C2PA Content Credentials specification, bind cryptographically verifiable assertions about the origin and editing history of media to the assets themselves [16]. The open research questions concern less the cryptography than the ecosystem: the robustness of soft bindings when credentials are stripped or content is laundered through re-encoding; the semantics of an *absent* credential in a partially adopted ecosystem; and, following Section 2.2, evaluation of provenance by its effect on attacker economics – the cost of making laundered content travel as far as authentic content – rather than as an artifact-level detection contest that scales poorly for defenders.

3.3 Attribution and Accountability

Between AI-generated content or autonomous systems and the actors responsible for creating, deploying, financing or amplifying them exists a chain of technical, organizational and institutional relationships that mediates attribution and determines accountability. These may include user accounts, authentication mechanisms, payment trails, hosting providers, model access pathways and distribution channels. This is also where fundamental rights are most exposed. The same mechanisms that can expose fraud networks, coordinated manipulation or criminal activity can also endanger journalists, human rights defenders or political dissidents if deployed without appropriate safeguards. Research on cybercrime and cyberterrorism has long emphasized that protection of fundamental rights must be designed into investigative and defensive capabilities rather than appended after deployment [17]. The objective is therefore not universal identification, but proportionate and contestable attribution.

3.4 Trust Tracking and Portable Reputation

Where identity establishes that an actor is singular and real, trust tracking lets positive and negative experience with that actor accumulate. The Montreal Memorandum envisions protocols with which individuals share their trust judgments across their networks, allowing reputation to spread without central authorities and without collapsing pseudonymity [15]. Open problems include resistance to reputation farming by fleets of AI agents, the diffusion of trust in adversarial networks, and the governance of infrastructures themselves.

3.5 Measuring Resilience and Harm

Cybersecurity matured to the extent that it was made measurable: vulnerability databases, incident response practices and disclosure norms created a shared empirical basis for defence. The digital space needs equivalents – harm observatories that monitor emerging manipulation tactics, botnets and synthetic influence operations, and resilience indicators that make the effect of interventions comparable [15]. Such indicators could feed directly into the data-driven benchmarking and monitoring frameworks for the SDGs, connecting SDG 16 reporting to information integrity.

4 From Agenda to Action

4.1 The Montreal Memorandum

The above is consistent with the outcome of recent multi-stakeholder deliberations. The 2025 Montreal Memorandum, developed by the community convened through the Digital Trust Convention, articulates the broader direction [15]. It diagnoses four crises – of trust, of attention, of markets and of common ground – driven by a flood of generative content and deceptively human-like bots, and sets out pathways for action across three interdependent strands: engineering systems based on available technologies, effective governance of their uses and impacts, and mobilisation of collective capacity to change entrenched incentives. It names *resilience-by-design* as an explicit principle and envisions a digital world that is “structurally resilient to the machinations of bad actors” [15].

4.2 A Converging Community

Restoring trust requires e.g. cryptographers and security economists alongside standardization experts, lawyers and regulators, platform engineers, behavioural and social scientists, journalists, humanitarian organizations and philanthropies. Since 2024, the Digital Trust Convention has convened this combination building a multistakeholder community for trust and resilience in the digital space [18].

4.3 An Incubation Pathway

We propose five steps. First, a shared taxonomy of threats and mechanisms. Second, pilots that deploy identity, provenance and trust-tracking mechanisms in bounded communities of practice – journalism, science communication, humanitarian response – where harms are acute. Third, evaluation of these

pilots against attacker economics rather than adoption alone. Fourth, governance capable of stewarding resilience mechanisms across jurisdictions through multistakeholder cooperation. Fifth, harm observatories and resilience indicators that link results to SDG monitoring. Participation from the Global South is essential from the outset: harms concentrate where resilience is thinnest. - Such an agenda is unlikely to progress unless there is an institutional catalyst that incubates initiatives and coalitions and holds the systemic map. The authors are creating the “Centre for a Positive Digital Future” to fill this role.

4.4 Governance Conditions and Safeguards

The proposed mechanisms can also create new risks. A resilient digital space must preserve the conditions that make digital participation open, inclusive and rights-respecting. Resilience mechanisms must address the risk of exclusion. Identity and personhood mechanisms must avoid disadvantaging individuals who lack access to reliable identification systems. Mechanisms designed to establish accountability must avoid creating new barriers to participation. Second, mechanisms capable of linking online activity to real-world actors must avoid misuse for political control, discrimination or suppression of expression. The objective should therefore not be universal traceability. Third, governance arrangements must anticipate institutional failure. Infrastructures themselves can become targets of manipulation, capture or abuse. Questions regarding who operates these systems and how errors are corrected are central research and governance questions. Cybersecurity practice has demonstrated that effective protection requires not only technical controls but also governance arrangements that define acceptable use, allocate responsibility and maintain trust. The same principle applies to the emerging infrastructures of the AI-saturated digital space.

5 Conclusion

AI governance is necessary but binds only those who agree to be bound. A substantial share of AI-induced risk is better understood as a property of the environment in which capable models act than of the models or their governable operators. Three decades of cybersecurity offer tested design principles for such environments – assume compromise, target attacker economics, allocate incentives deliberately, prefer protocols over user burden – and the mechanisms of identity, provenance, attribution, trust tracking and accountability translate these principles into a concrete research agenda. The Montreal Memorandum provides the direction, and the Digital Trust Convention the community for pursuing this agenda. Within this broader approach, AI governance still matters but becomes a narrower and considerably more tractable task. Upgrading the infrastructure of the digital space as a whole is meanwhile the sustainable response to the bad actor problem, and it directly serves resilient infrastructure (SDG 9), peaceful, just and inclusive institutions (SDG 16) and the partnerships (SDG 17) through which the SDGs are to be achieved.

References

- [1] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 (Artificial Intelligence Act), Official Journal of the European Union, L series, 2024.
- [2] National Institute of Standards and Technology, “Artificial Intelligence Risk Management Framework (AI RMF 1.0),” NIST AI 100-1, Gaithersburg, MD, USA, Jan. 2023.
- [3] M. Brundage et al., “The malicious use of artificial intelligence: Forecasting, prevention, and mitigation,” arXiv:1802.07228, 2018.
- [4] V. Ciancaglini, C. Gibson, D. Sancho, P. Amann, A. Klayn, O. McCarthy, and M. Eira, “Malicious uses and abuses of artificial intelligence,” Trend Micro Research, UNICRI and Europol EC3, Nov. 2020.
- [5] X. Qi, Y. Zeng, T. Xie, P.-Y. Chen, R. Jia, P. Mittal, and P. Henderson, “Fine-tuning aligned language models compromises safety, even when users do not intend to!,” in Proc. Int. Conf. Learning Representations (ICLR), 2024.
- [6] S. Lermen, C. Rogers-Smith, and J. Ladish, “LoRA fine-tuning efficiently undoes safety training in Llama 2-Chat 70B,” arXiv:2310.20624, 2023.
- [7] United Nations, “Global Digital Compact,” annexed to the Pact for the Future, UN General Assembly, Sep. 2024.
- [8] S. Rose, O. Borchert, S. Mitchell, and S. Connelly, “Zero trust architecture,” NIST Special Publication 800-207, Gaithersburg, MD, USA, Aug. 2020.
- [9] R. Anderson and T. Moore, “The economics of information security,” Science, vol. 314, no. 5799, pp. 610–613, 2006.
- [10] CyberPeace Institute, “Playing with lives: Cyberattacks on healthcare are attacks on people,” Geneva, Switzerland, Mar. 2021.
- [11] C. Herley, “So long, and no thanks for the externalities: The rational rejection of security advice by users,” in Proc. New Security Paradigms Workshop (NSPW), 2009, pp. 133–144.
- [12] A. P. Felt, R. Barnes, A. King, C. Palmer, C. Bentzel, and P. Tabriz, “Measuring HTTPS adoption on the web,” in Proc. 26th USENIX Security Symp., 2017, pp. 1323–1338.
- [13] S. Adler et al., “Personhood credentials: Artificial intelligence and the value of privacy-preserving tools to distinguish who is real online,” arXiv:2408.07892, 2024.
- [14] Regulation (EU) 2024/1183 of the European Parliament and of the Council of 11 April 2024 amending Regulation (EU) No 910/2014 Framework, Official Journal of the European Union, L series, 2024.
- [15] Digital Trust Convention, “The Montreal Memorandum,” 2025. [Online]. Available: <https://www.montrealmemorandum.org>
- [16] Coalition for Content Provenance and Authenticity, “Content Credentials: C2PA technical specification,” version 2.3, Jan. 2026. [Online]. Available: <https://spec.c2pa.org>
- [17] F. Bosco, E. Vermeersch, V. Luda, G. Vaciago, U. Gasper, and A. Lyle, “Non-discrimination and protection of fundamental rights in cybercrime and cyberterrorism research,” in Combatting Cybercrime and Cyberterrorism: Challenges, Trends and Priorities, B. Akhgar and B. Brewster, Eds. Cham, Switzerland: Springer, 2016, pp. 97–115.
- [18] OECD.AI Policy Observatory, “The Digital Trust Convention.” [Online]. Available: <https://oecd.ai/en/digital-trust>

AI-Based Analysis of Online Data for Climate Hazard Resilience: From Digital Traces to SDG 13 Indicators

Dmitry Erokhin, PhD*
International Institute for Applied
Systems Analysis
Laxenburg, Austria
erokhin@iiasa.ac.at

Nadejda Komendantova, PhD*
International Institute for Applied
Systems Analysis
Laxenburg, Austria
komendan@iiasa.ac.at

Abstract

During heatwaves, floods, and wildfires, people seek immediate advice and discuss the credibility of official warnings online. Search data, posts, and comments can reveal changing communication needs alongside meteorological and impact records. This study examines how AI-supported analysis of these traces could inform climate-risk communication under Sustainable Development Goal 13. The proposed procedure combines natural language processing, topic and emotion analysis, event-time comparisons, and expert review to derive five candidate indicators. Three tests assess classification accuracy, agreement with independent evidence, and usefulness to emergency communication teams. Each measure requires a clear denominator, geographical and linguistic coverage, uncertainty estimates, and privacy safeguards. Official statistics, surveys, and community engagement provide context for interpreting online patterns. Prospective testing across hazards and settings would show where these measures provide reliable guidance.

Keywords

AI for SDGs, SDG 13, climate resilience, digital traces, risk communication, responsible AI

1 Introduction

Climate change is increasing the frequency or intensity of many climate-related hazards [1]. As an event approaches, public questions become increasingly practical. Heatwaves prompt searches for symptoms and safe ways to cool down. During a flood, attention may turn to road closures, evacuation routes, and confidence in official warnings. The timing of these questions may reveal gaps in risk communication.

These evolving information needs relate to SDG 13 and its commitments to resilience, adaptive capacity, education, and awareness [2]. Official SDG indicators offer comparable statistics, while online activity can reveal immediate questions raised by particular events [3]. This study asks when AI-derived digital traces could provide reliable evidence for risk communication. Five candidate measures are defined alongside

tests of their validity and usefulness. Research on AI and sustainable development informs the approach [4, 5].

Earlier studies provide an empirical basis for this question. Google Trends can indicate changing information needs [6], and social media research has examined disaster awareness and damage assessment [7, 8]. Studies of climate narratives and hazard resilience have extended these approaches to climate-related concerns [9, 10]. Research on disaster misinformation highlights the importance of contextual verification [11, 12]. The Galicia study combined search behaviour and discussions during climate hazards [10]. Figure 1 brings these strands together in a procedure for deriving, interpreting, and reviewing comparable candidate indicators.

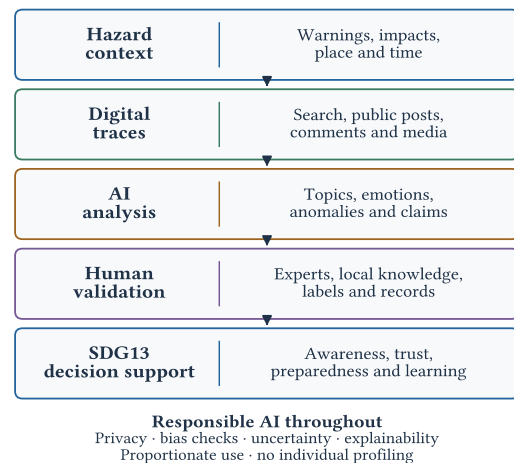


Figure 1. Procedure for deriving and reviewing candidate indicators.

2 Sources and Candidate Indicators

Figure 1 links the online analysis to the physical event. Meteorological observations, official warnings, exposure and impact records, emergency timelines, and media coverage establish the timing and severity of each hazard. Search-interest

* The authors contributed equally. Correspondence: erokhin@iiasa.ac.at.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Information Society 2026, Ljubljana, Slovenia

© 2026 Copyright held by the owner/author(s).

DOI: <https://doi.org/10.70314/is.2026.ai4sdg.143>

series can then show changes in information demand, while publicly available posts and comments add evidence on questions, narratives, emotional responses, and views of public institutions. Data collection should follow the ethical and access requirements of each platform.

With the event context established, the analysis uses aggregated online data. Natural language processing identifies locations, protective actions, institutions, and factual claims. Topic modelling groups related discussions, while sentiment and emotion analysis adds information about tone. Anomaly detection highlights departures from a documented pre-event baseline. Expert review addresses ambiguous items and model errors using independent records and local knowledge. The reporting record preserves collection rules, denominators, and uncertainty estimates.

Table 1 translates these analytical outputs into five candidate measures linked to SDG targets 13.1 on resilience and adaptive capacity and 13.3 on education and awareness [2, 3]. Each measure requires a documented source, geographical scope, observation period, denominator, validation procedure, and uncertainty statement. For example, the share of flood-related comments asking about safe routes could be compared across the warning and impact periods, with road-closure records providing the event context. Interpreting that share would also require information about the platform's local coverage and the volume of eligible comments. Such checks connect the candidate measures to specific communication questions and provide a starting point for independent validation.

Table 1. Candidate indicators and safeguards.

Indicator	Candidate measure	Interpretation safeguard
Public sensitivity	Search interest against a pre-event baseline	Check hazard severity and media coverage
Preparedness discourse	Share of items mentioning protective action	Verify protective actions with independent evidence
Misinformation pressure	Share with human-verified false or misleading claims	Report unverified model flags separately
Institutional trust	Supportive and distrustful shares of institution-related items	Assess factual claims and institutional criticism separately
Digital inclusion	Coverage by place, language, and platform	Describe coverage limits and seek independent demographic data

3 Event-Time and Hazard-Specific Interpretation

The meaning of an online signal also changes as a hazard unfolds. Figure 2 illustrates possible trajectories before, during, and after an event. Early searches for protective action may reflect preparations following a warning. The same terms during impact can indicate urgent needs, uncertainty about

instructions, or renewed media attention. Continued interest during recovery may reveal unanswered questions about services and assistance, although some activity may come from observers outside the affected area. Comparing online series with warning times, impact reports, and survey evidence would clarify these competing explanations. The curves in Figure 2 are illustrative and establish the comparisons an empirical study would need to make.

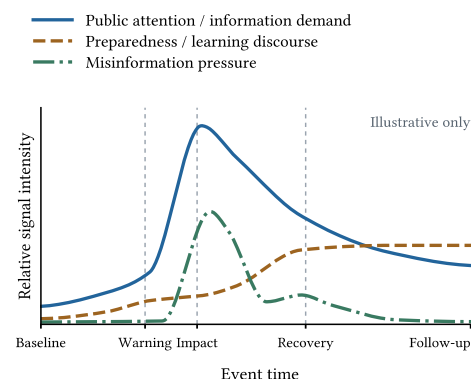


Figure 2. Illustrative event-time patterns. Curves are schematic.

The relevant questions also vary by hazard. Heatwave discussions may concern symptoms, cooling facilities, outdoor workers, and older people. Flood-related posts often focus on roads, evacuation, river levels, and emergency assistance. A study of wildfire discussions identified emotional responses, criticism of governance, and conspiracy narratives [13]. Questions about drought restrictions, agriculture, prices, and fairness may develop over longer periods. Claim verification should therefore reflect the characteristics of each hazard, with institutional criticism recorded separately from factual accuracy.

4 Implementation and Validation

An empirical application can test these interpretations by comparing hazards across predefined warning, impact, and recovery periods. The research record should document search terms, retrieval dates, API or export settings, eligibility rules, deduplication, language and location criteria, model versions, prompts or classifier settings, random seeds where relevant, and aggregation procedures. Each content share would use unique eligible items from the same platform, language, place, and period as its denominator. Search indices and post counts would be presented separately because they represent different quantities. A shared event log would record hazard severity, official messages, and major media peaks.

The design also requires careful attention to platform differences. An event-centred El Niño study analysed 8,145 relevant items across six platforms and found differences in language, topics, and sentiment [14]. These differences support platform-specific evaluation before signals are compared. Claims about preparedness or warning performance require additional evidence from surveys, operational records, or observed disaster outcomes.

The first evaluation stage examines classification accuracy. Two independent annotators would label a held-out sample stratified by language, platform, hazard, and event phase using

a documented codebook and adjudication procedure. The study would report stratum sizes, inter-annotator agreement, class-specific precision and recall, macro-F1, calibration, and uncertainty intervals. Separate tuning and test episodes, together with deduplication across splits, would reduce leakage. Suspected misinformation would enter an indicator only after human verification.

The second stage examines indicator validity by comparing aggregate online measures with independent warning and impact records. Surveys or participatory evidence would add information for preparedness and trust. Sensitivity analyses would vary queries, pre-event baselines, aggregation windows, platform inclusion, and classification thresholds. Repeated testing across hazards would show how consistently a measure transfers between settings. Sparse coverage would be recorded explicitly because low activity can reflect limited digital access, while large spikes can arise from distant media attention.

With validity established, the third stage assesses usefulness to communication teams. Candidate measures would be compared with simple keyword and volume counts and with existing monitoring procedures. Evaluation would cover detection lead time at comparable false-alert rates, confirmed information gaps, sensitivity to analytical choices, and practitioner assessments. Decision thresholds would be agreed before deployment and documented after each event. A prospective pilot would examine the timeliness, clarity, and practical value of the resulting briefings.

The practical evaluation can also motivate research on the effects of communication. Questions about whether communication changes preparedness call for a separate causal study. An interrupted time series would need a clearly timed intervention and controls for seasonality, autocorrelation, and concurrent events. A difference-in-differences design would require credible parallel trends, limited anticipation, and attention to spillovers. When these assumptions remain uncertain, interpretation should stay with observed patterns and associations.

5 Discussion, Responsible AI and Limitations

Figure 3 connects the validation stages to a reporting cycle for communication teams. Historical data may reveal recurring questions as agencies prepare warnings. During an emergency, short briefings could highlight urgent information needs and claims awaiting verification. After the event, practitioners could review persistent questions and adjust future messages. This sequence reflects the emphasis on people-centred warning systems in the Sendai Framework and Early Warnings for All [15, 16].

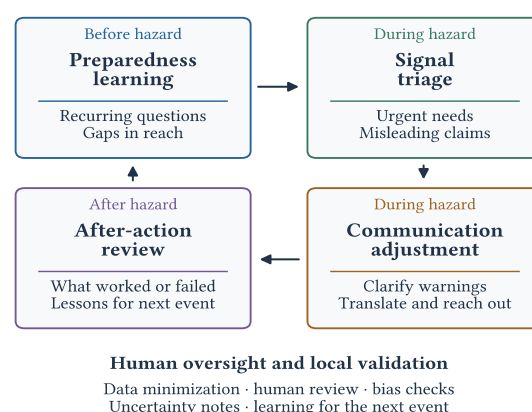


Figure 3. Reporting and learning with human oversight.

For this cycle to reflect conditions on the ground, the interpretation of online evidence must involve people familiar with the affected communities. Participatory adaptation research highlights the contribution of local expertise and stakeholder engagement [17, 18]. Studies of social media use for disaster risk reduction identify perceived usefulness, platform characteristics, and source credibility as relevant factors [19]. Research on misinformation governance and AI tools further emphasises context and institutional responsibility [20, 21]. A classifier may perform well while the underlying material remains concentrated among frequent users or residents of better-connected areas. Reviewing the findings with local organisations and communication staff can reveal these gaps and help distinguish online attention from unmet needs. Public messages and emergency decisions would remain the responsibility of the relevant agencies.

Ethical and operational safeguards follow the same logic. OECD and UNESCO guidance emphasises transparency, fairness, accountability, and human oversight [22, 23]. The proposed analysis uses aggregated data, limits collection to necessary information, and documents retention and deletion procedures. Language models can misinterpret irony, multilingual expressions, and criticism, while sentiment analysis can reproduce gender bias [24]. Differences in internet access and platform visibility affect geographical coverage, and changing APIs complicate replication. Community consultation is especially valuable where online representation is limited. A transparent classifier or simple anomaly detector may be suitable when its performance can be checked directly.

6 Conclusion

The five candidate measures give this study a basis for evaluating the use of online activity in climate-risk communication. Interpretation rests on clear denominators, event records, independent comparisons, and human review. Surveys, participatory research, and impact records add evidence about preparedness and resilience. A preregistered prospective study across several hazards and settings could establish which measures remain reliable under changing conditions and whether their use improves risk communication.

An initial pilot could involve a local communication team that records which online questions inform changes in official messages and how quickly each signal is verified. Documenting the languages and locations with weak digital coverage would help direct complementary community interviews. Subsequent events would provide evidence on whether the indicators remain useful under different hazard and communication conditions.

References

- [1] IPCC. 2023. *Climate Change 2023: Synthesis Report*. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland, pp. 35–115. <https://doi.org/10.59327/IPCC/AR6-9789291691647>.
- [2] United Nations. n.d. *Goal 13: Take urgent action to combat climate change and its impacts*. <https://sdgs.un.org/goals/goal13>. Accessed 23 September 2026.
- [3] United Nations Statistics Division. n.d. *SDG Indicators Metadata Repository, Goal 13*. <https://unstats.un.org/sdgs/metadata/?Goal=13>. Accessed 23 September 2026.
- [4] David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, et al. 2022. Tackling climate change with machine learning. *ACM Computing Surveys* 55, 2, Article 42, 96 pages. <https://doi.org/10.1145/3485128>.
- [5] Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, et al. 2020. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications* 11, 233. <https://doi.org/10.1038/s41467-019-14108-y>.
- [6] Amaryllis Mavragani and Gabriela Ochoa. 2019. Google Trends in infodemiology and infoveillance: methodology framework. *JMIR Public Health and Surveillance* 5, 2, e13439. <https://doi.org/10.2196/13439>.
- [7] Dmitry Erokhin and Nadejda Komendantova. 2024. Social media data for disaster risk management and research. *International Journal of Disaster Risk Reduction* 114, 104980. <https://doi.org/10.1016/j.ijdr.2024.104980>.
- [8] Yury Kryvasheyev, Haohui Chen, Nick Obradovich, Esteban Moro, Pascal Van Hentenryck, James Fowler, et al. 2016. Rapid assessment of disaster damage using social media activity. *Science Advances* 2, 3, e1500779. <https://doi.org/10.1126/sciadv.1500779>.
- [9] Dmitry Erokhin and Nadejda Komendantova. 2024. Unveiling the dynamics of climate change narratives: A Google Trends analysis. *Observatorio (OBS*)* 18 (Special Issue), 22–42. <https://doi.org/10.15847/obsOBS18520242567>.
- [10] Dmitry Erokhin and Nadejda Komendantova. 2026. Assessing Climate Hazard Resilience Through AI-Based Analysis of Online Data: Empirical Evidence from Galicia. *Societies* 16, 6, 188. <https://doi.org/10.3390/soc16060188>.
- [11] Irina Dallo, Or Elroy, Laure Fallou, Nadejda Komendantova, and Abraham Yosipof. 2023. Dynamics and characteristics of misinformation related to earthquake predictions on Twitter. *Scientific Reports* 13, 13391. <https://doi.org/10.1038/s41598-023-40399-9>.
- [12] Nadejda Komendantova and Dmitry Erokhin. 2025. Artificial intelligence tools in misinformation management during natural disasters. *Public Organization Review* 25, 1, 81–105. <https://doi.org/10.1007/s11115-025-00815-2>.
- [13] Dmitry Erokhin. 2025. Public Discourse Surrounding the 2025 California Wildfires: A Sentiment and Topic Analysis of High-Engagement YouTube Comments. *Geosciences* 15, 3, 100. <https://doi.org/10.3390/geosciences15030100>.
- [14] Dmitry Erokhin and Nadejda Komendantova. 2026. El Niño Discourse and the Limits of Single-Platform Inference. *Information* 17, 7, 622. <https://doi.org/10.3390/info17070622>.
- [15] United Nations Office for Disaster Risk Reduction (UNDRR). 2015. *Sendai Framework for Disaster Risk Reduction 2015–2030*. <https://www.undrr.org/publication/sendai-framework-disaster-risk-reduction-2015-2030>. Accessed 23 September 2026.
- [16] United Nations Office for Disaster Risk Reduction (UNDRR). n.d. *Early Warnings for All*. <https://www.undrr.org/early-warnings-for-all>. Accessed 23 September 2026.
- [17] Enora Bruley, Anna Scolobig, Marta Ellena, Sam Pickard, Mathilda Englund, Eulàlia Baulenas, et al. 2026. Advancing stakeholder engagement in climate adaptation: a systematic review of structural barriers and operational enablers. *Sustainability Science*. <https://doi.org/10.1007/s11625-026-01843-7>.
- [18] Nadejda Komendantova, Dmitry Erokhin, Anna Scolobig, Enora Bruley, Marina Mattera, and Madeline Baldelli. 2026. Co-creating adaptation solutions: A critical review of participatory instruments in climate change adaptation laws and policies. *Earth System Governance* 28, 100326. <https://doi.org/10.1016/j.esg.2026.100326>.
- [19] Tahereh Zobeidi, Nadejda Komendantova, Masoud Yazdanpanah, and Alexa Lamm. 2024. A multi-dimensional model of anticipating intention to use social media for disaster risk reduction. *International Journal of Disaster Risk Reduction* 104, 104356. <https://doi.org/10.1016/j.ijdr.2024.104356>.
- [20] Nadejda Komendantova, Love Ekenberg, Mattias Svahn, Aron Larsson, Syed Iftikhar Hussain Shah, Myrsini Glinos, et al. 2021. A value-driven approach to addressing misinformation in social media. *Humanities and Social Sciences Communications* 8, 33. <https://doi.org/10.1057/s41599-020-00702-9>.
- [21] Rosa Vicari and Nadejda Komendantova. 2023. Systematic meta-analysis of research on AI tools to deal with misinformation on social media during natural and anthropogenic hazards and disasters. *Humanities and Social Sciences Communications* 10, 332. <https://doi.org/10.1057/s41599-023-01838-0>.
- [22] OECD. 2024. *OECD AI Principles*. Adopted 2019, updated 2024. OECD, Paris. <https://www.oecd.org/en/topics/ai-principles.html>. Accessed 23 September 2026.
- [23] UNESCO. 2022. *Recommendation on the Ethics of Artificial Intelligence*. Adopted 23 November 2021. UNESCO, Paris. <https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence-0>. Accessed 23 September 2026.
- [24] Mohammad Reza Yeganegi, Hossein Hassani, and Nadejda Komendantova. 2025. Identifying and Mitigating Gender Bias in Social Media Sentiment Analysis: A Post-Training Approach on Example of the 2023 Morocco Earthquake. *Information* 16, 8, 679. <https://doi.org/10.3390/info16080679>.

From Documents to Signals: Toward an AI-powered Policy Intelligence Framework

Masato Oda
OSINTech Inc.,
Kobe Institute of Computing
Kobe Japan
masato@osintech.net

Masamichi Kobayashi
OSINTech Inc.,
Graduate School of Societal Safety
Sciences, Kansai University
k.masamichi@osintech.net

Kazue Oda
OSINTech Inc.
kazue@osintech.net

Tomoyuki Fuse
OSINTech Inc.
tomoyuki@osintech.net

Sho Taya
OSINTech Inc.
taya@osintech.net

Abstract

Conventional policy research has struggled to identify weak and dispersed policy concerns systematically and at scale when they emerge across countries, languages, institutions, and policy domains. This study examines whether an AI-assisted approach can identify policy trends using primary policy sources collected and structured across national, linguistic, and organizational boundaries.

We used RuleWatcher, which covers multilingual primary policy sources published by governments in approximately 150 countries, as well as communications from organizations involved in rulemaking, including United Nations bodies and non-governmental organizations (NGOs). For the empirical analysis, we constructed a dataset of 2,753 records concerning AI ethics.

We divided the data into an earlier period (1 January 2019 to 31 December 2024) and a later period (1 January 2025 to 31 July 2026). From expressions found in the earlier-period documents, we extracted 132 keywords after consolidating synonymous terms. We then used AI-based semantic similarity assessment and author review to organize these keywords into 26 "Concept Cards." We counted the occurrence of expressions associated with each Concept Card by policy actor and quarter, and examined whether concepts detected in the earlier period could be treated as early indications of policy agenda formation when their frequency increased in the later period.

Of the 2,753 records analyzed, 1,116 contained expressions associated with at least one Concept Card. Counting documents assigned to multiple concepts in each applicable category produced 2,671 concept-document assignments. Of the assignments observed in the earlier period, 65.6% came from non-governmental actors, including NGOs. The quarterly distribution also showed that references by non-governmental actors continued from the early part of the observation period. Across complete quarters, excluding the partial third quarter of

2026, the mean quarterly count for all actors increased from 77.3 in the earlier period to 127.3 in the later period. Comparison of the 26 concepts by actor and quarter revealed a temporal difference for multiple concepts: early references by non-governmental actors were followed by increased references from UN bodies and governments.

These findings suggest that AI-based classification and temporal analysis of multilingual, multi-actor primary policy sources can provide evidence for observing the formation of policy agendas. However, the observed temporal differences do not demonstrate causal diffusion between actors or policy adoption.

Based on these findings, this study proposes a "Governance Observatory" as an AI-assisted intelligence framework for continuously observing policy agendas as they emerge.

Humanity increasingly faces challenges that no single country can address alone. Perspectives dispersed across underrepresented regions and linguistic minorities may contribute to addressing such challenges. Yet the rapid adoption of AI in policymaking may introduce centralization bias and further marginalize communications from these regions and communities. To mitigate this risk, policymakers should consider using AI to monitor a broad range of primary policy sources continuously, including communications from non-governmental actors.

Keywords

AI-assisted policy analysis, policy agenda formation, Governance Observatory, horizon scanning, RuleWatcher

1. Background and Research Objectives

Policy is one of the principal means by which public institutions address societal challenges. However, existing policy frameworks may not respond adequately to complex global problems such as climate change and social instability caused by inequality. This limitation is particularly pronounced for rapidly advancing technologies such as AI: regulation and rulemaking often lag behind social change, while concerns that are currently weak and dispersed but may become important to future policymaking or the resolution of global challenges remain insufficiently recognized.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, 5–9 October 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
DOI : <https://doi.org/10.70314/is.2026.ai4sdg.157>

To respond to increasingly complex challenges in advance, policymakers need systematic ways to identify concerns that have not yet been clearly articulated as policy issues but are appearing sporadically, and to present them in a form that can inform policy decisions.

Hajer [1] and Ingram et al. [2] have shown how policy issues are formed through discourse and social meaning-making. Building on this perspective, this study focuses on dispersed references that appear before an issue becomes clearly established on the policy agenda.

Practical approaches to future-oriented policy analysis include strategic foresight [3], scenario planning [4], and horizon scanning. Horizon scanning, in particular, is regarded as useful for science, technology, and innovation (STI) policy and governance because it identifies "weak signals," functions as an early-warning system for emerging issues, and provides insights to policymakers [5].

Conventional horizon-scanning exercises often select observation targets through expert nomination or predefined information sources and classification frameworks. Depending on the research design, such approaches may overlook early concerns that do not fit established problem definitions or that originate from peripheral actors and languages.

This study uses AI to address a question that remains insufficiently examined: how can perceptions of policy problems dispersed across policy domains be observed before they are explicitly recognized as policy issues? We report the results of this investigation.

2. Research Questions

This study examines a general analytical approach for identifying policy agendas as they emerge from multilingual primary policy sources published by diverse actors. AI ethics is used as an empirical case to assess the feasibility of this approach. Accordingly, we address the following research questions:

RQ1. To what extent can AI-assisted conceptual classification and temporal analysis of multilingual primary policy sources reveal the early formation of policy agendas?

RQ2. When and how consistently do policy agendas appear in primary policy sources published by non-governmental actors, including NGOs and specialist organizations, UN bodies, and national governments? How do their frequency and temporal trajectories differ across policy actors?

RQ3. How can AI-based candidate detection be combined with human expert review to continuously identify indications of emerging policy agendas in multilingual primary policy sources published by diverse actors?

3. Method

We used RuleWatcher [6], which standardizes multilingual primary policy sources into a common analytical language and enables comparison across countries, regions, and organization types. Taking AI ethics as an empirical case, we assessed an analytical method for identifying early policy agenda formation and considered its broader applicability.

Overview of the RuleWatcher Data

Number of records: 1,742,102

Coverage: news releases from governments in 151 countries, 59 UN bodies, 904 NGOs and other non-governmental actors, and other sources

Publication period: 1 January 2019 to 31 July 2026

Multilingual primary policy sources standardized into English

17 themes and approximately 1,500 topic tags

The AI Governance theme uses a search framework based on UNESCO's Recommendation on the Ethics of Artificial Intelligence [7]

Dataset Construction

From 1,742,102 RuleWatcher records (a), we extracted 96,751 records (b) published by governments, UN bodies, and non-governmental actors and classified under the AI Governance theme.

From dataset (b), we extracted records assigned the predefined RuleWatcher topic tag "AI Ethics." After duplicate and relevance screening, 2,753 records (c) remained in the empirical dataset.

We classified dataset (c) by organization type into government (d), UN bodies (e), and other non-governmental actors (f).

For each organization type, we divided the records into the earlier period, 1 January 2019 to 31 December 2024 (g, h, i), and the later period, 1 January 2025 to 31 July 2026 (j, k, l).

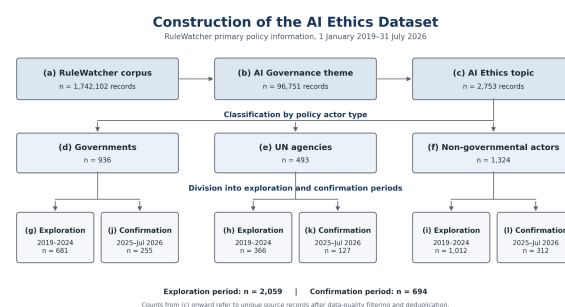


Figure1: Dataset construction

4. Data Analysis

We divided corpora (g)–(l) into quarterly subsets and conducted a two-stage analysis using OpenAI's GPT-5.6-sol.

RuleWatcher retained the original-language records and standardized their textual content into English through its machine-translation pipeline. The present analysis was conducted on the standardized English descriptions. GPT-5.6-sol was applied using a fixed prompt across all quarterly subsets. Because generative-model outputs may vary between runs and model versions, the final list of 132 expressions and their mappings to the 26 Concept Cards were preserved as fixed analytical artifacts.

In the first stage, we extracted candidate expressions from the earlier-period data using AI" or "we used AI to extract candidate expressions from the earlier-period data. Expressions appearing at least twice within a document were evaluated using the

following four criteria. After removing duplicates from the 500 highest-ranked expressions, 132 expressions were retained.

- Novelty: expressions rarely used in the past
- Growth: expressions whose frequency increased from the previous quarter
- Burst: expressions showing an anomalous increase over a short period
- TF-IDF: expressions frequent within individual documents but not widely distributed across the full dataset

The 132 expressions included synonymous and related terms that differed in wording but referred to a common problem concept. For example:

- AI environmental impact
- energy consumption of AI
- carbon footprint of AI
- data center emissions
- environmental cost of foundation models

Although phrased differently, these expressions can all be understood as referring to the common problem concept of the environmental impact of AI. We therefore combined AI-based semantic similarity assessment with author review to organize the 132 expressions into 26 semantic groups, which we define as "Concept Cards."

In the second stage, we mapped the 132 expressions to their corresponding Concept Cards and identified documents containing at least one expression associated with each card. We then counted documents by Concept Card, policy actor, and quarter.

Within each combination of Concept Card, policy actor, and quarter, each document was counted only once. A document associated with multiple Concept Cards was counted once under each applicable card.

This analysis examined whether the frequency of problem concepts detected in the earlier period increased in the later period, and whether their timing, persistence, and temporal trajectories differed across policy actors.

Reproducibility and Validation Protocol

To ensure traceability, the analytical workflow was documented at each stage, including corpus extraction, quarterly segmentation, candidate-expression ranking, semantic consolidation, document matching, and aggregation. The reproducibility materials specify the dataset inclusion and exclusion criteria, the 132 candidate expressions and their mappings to the 26 Concept Cards, the prompts used for AI-assisted classification, the model configuration, and the rules used to count documents by concept, actor, and quarter. The Concept Card dictionary was fixed before its application to the later-period data, thereby preventing later observations from changing the classification framework.

AI-generated groupings were not accepted automatically. The authors reviewed whether the expressions assigned to each Concept Card referred to the same or a closely related policy

problem, revised ambiguous groupings, and confirmed the final mappings. Exact keyword matching was then applied to the standardized English descriptions. Within each Concept Card, policy actor, and quarter, a document was counted only once, although a document could be assigned to multiple Concept Cards.

5. Results

Applying the 26 Concept Cards developed from the earlier-period data, we detected keywords associated with at least one card in 1,116 of the 2,753 documents (40.5%). Counting documents assigned to multiple concepts under each applicable card yielded 2,671 concept-document assignments.

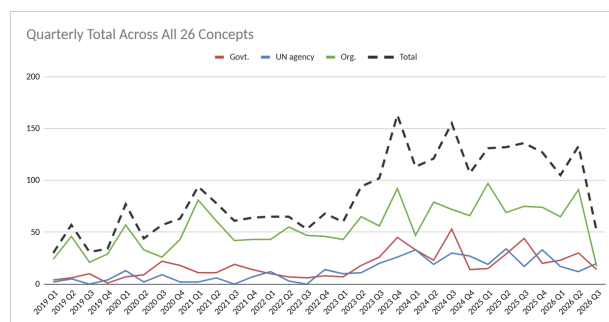


Figure 2: Quarterly concept-document assignments by policy actor across 26 concepts

Trends in the Number of Documents. For 2026, the data cover only January–July.

Across the full observation period, non-governmental actors accounted for 1,705 concept-document assignments, governments for 557, and UN bodies for 409 (Figure 2). Excluding the partial third quarter of 2026, the mean quarterly count across all actors increased by 64.7%, from 77.3 in the earlier period to 127.3 in the later period. Counts increased for all three actor groups; the mean for UN bodies approximately doubled, from 10.7 to 22.0.

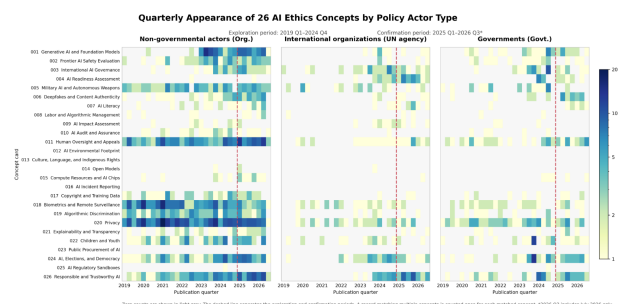


Figure 3: Heatmap of 26 concepts by policy actor and quarter

Figure 3 presents the frequency of the 26 concepts by policy actor and quarter. Aggregated across all actors, five concepts - Privacy; Human Oversight and Appeals; Biometrics and Remote Surveillance; Children and Youth; and AI, Elections, and Democracy - appeared in all 31 observed quarters. Responsible and Trustworthy AI and Algorithmic Discrimination appeared in

30 of the 31 quarters. These patterns indicate that the concepts were addressed consistently throughout the observation period rather than appearing as temporary topics.

By contrast, Generative AI and Foundation Models, Deepfakes and Content Authenticity, AI Readiness Assessment, and AI Literacy increased in frequency particularly after 2023. The AI ethics dataset thus contains both long-standing concerns and concerns that became more visible in association with recent technological and institutional changes.

We next compared the first quarter in which each concept was detected within the dataset across policy actors. Of the 21 concepts observed for both non-governmental actors and UN bodies, detection by non-governmental actors occurred earlier for 17 concepts, in the same quarter for three, and later for one. Of the 21 concepts observed for both non-governmental actors and governments, detection by non-governmental actors occurred earlier for 15 concepts, in the same quarter for five, and later for one.

These findings do not mean that non-governmental actors preceded other actors for every problem concept. Nevertheless, among the 21 comparable concepts, non-governmental actors were detected earlier for 17 concepts relative to UN bodies and for 15 relative to governments. Multiple problem concepts were therefore observable in the primary policy sources of non-governmental actors at a comparatively early stage.

Here, "first detection" refers to the first quarter within the dataset's observation period in which at least one keyword associated with a Concept Card was detected. It does not identify when a problem concept was first proposed in practice, when it was institutionally adopted, or whether one policy actor directly influenced another. The results nevertheless indicate that continuous observation of primary policy sources from diverse actors may identify problem concepts before they become widely recognized as policy agendas.

6. Conclusion

This study examined whether AI-based candidate detection, conceptual classification, and temporal analysis of multilingual primary policy sources published by diverse actors could identify early indications of emerging policy agendas. In the empirical case of AI ethics, differently worded concerns were organized into common conceptual units, enabling their timing, frequency, and persistence to be visualized across actors. The findings demonstrate the potential of primary policy sources for observing temporal patterns in policy agenda formation.

Non-governmental actors, including NGOs, accounted for 65.6% of the concept-document assignments observed in the earlier period. Many problem concepts were detected in their primary policy sources before being detected in those of UN bodies or governments, after which counts increased across all actors. This temporal ordering does not demonstrate direct influence between actors or policy adoption. The findings may also be affected by differences in publication frequency, digital publishing infrastructure, RuleWatcher's coverage, machine-translation accuracy, and communication strategies across policy actors. Consequently, the observed counts do not directly measure the societal importance or actual prevalence of each problem concept.

Proposal: A Governance Observatory to Support Policymaking

Based on these findings, this study proposes the "Governance Observatory," an intelligence framework that continuously identifies indications of emerging policy agendas and presents them in a form that can inform policymakers' decisions.

As the use of AI in policy research expands, bias in the information selected for analysis may reproduce and amplify dominant regional, linguistic, and institutional values. The Governance Observatory therefore encompasses primary policy sources not only from governments and major international organizations, but also from NGOs, specialist organizations, actors in underrepresented regions, and linguistic minorities. By presenting weak and dispersed concerns in a form that can be examined, the framework aims to deepen understanding of complex policy challenges and provide a foundation for more inclusive and well-grounded policymaking.

Future research should assess the reproducibility and generalizability of the framework by applying it to policy fields beyond AI ethics.

Data and Materials Availability

RuleWatcher is a proprietary platform, and the record-level corpus used in this study cannot be publicly redistributed because it contains materials collected from external primary sources. The authors have retained the corpus-selection criteria, the 132-expression dictionary, the mappings to the 26 Concept Cards, the AI prompt and model configuration, the validation protocol, and the aggregated quarterly counts underlying Figures 2 and 3. Requests for access to these derived research materials for scholarly verification may be directed to the corresponding author and will be considered subject to applicable third-party rights and platform terms. Information about RuleWatcher and its access options is available at <https://web.rulewatcher.com/>.

References / Literatura

- [1] Maarten A. Hajer. 1995. *The Politics of Environmental Discourse: Ecological Modernization and the Policy Process*. Oxford University Press, Oxford, UK.
- [2] Helen Ingram, Anne L. Schneider, and Peter deLeon. 2007. *Social Construction and Policy Design*. In Paul A. Sabatier (Ed.), *Theories of the Policy Process*, 2nd ed., 93–126. Westview Press, Boulder, CO.
- [3] OECD. 2019. *Strategic Foresight for Better Policies: Building Effective Governance in the Face of Uncertain Futures*. OECD Publishing, Paris, France.
- [4] Kees van der Heijden. 2005. *Scenarios: The Art of Strategic Conversation*, 2nd ed. John Wiley & Sons, Chichester, UK.
- [5] OECD. 2026. *Building Capacity for Technology Horizon Scanning: A Guide for Policymakers*. OECD Science, Technology and Industry Working Papers, No. 2026/06. OECD Publishing, Paris, France.
- [6] OSINTech Inc. 2026. *RuleWatcher: A global policy data tool*. It offers extensive coverage of developing countries and includes a wealth of information from non-governmental entities. By providing local-language content in English, it directs users to primary sources. Users can also download structured data categorized by over 1,600 expert-selected tags. Access at: <https://web.rulewatcher.com/> (Accessed September 29, 2026).
- [7] UNESCO. 2022. *Recommendation on the Ethics of Artificial Intelligence*. United Nations Educational, Scientific and Cultural Organization, Paris, France.

Documented, Not Governed: Public Algorithm Registers and the Runtime Gap in EU AI Act Deployer Obligations

Zoran Nasteski
zoran@nasteski.com
QX
Florence, Italy

Teodor Bogoeski
teodorbogoeski@gmail.com
QX
Florence, Italy

Hristijan Jankulovski
jankulovski.hristijan@gmail.com
QX
Florence, Italy

Abstract

European public bodies publish algorithm registers describing the automated systems they operate. The Dutch national register, among the largest in Europe, additionally publishes a risk class for every entry, one of which corresponds to the EU AI Act's high-risk category and therefore signals legal exposure. We audit that register at three levels: its publication schema, all 1,534 published entries, and its complete revision history, comprising 2,921 published versions of 1,556 systems, 638 of which were revised. Of the 111 system names published by more than one body, 43 carry more than one class, and in eight of these, covering 91 entries, some bodies declare a system high risk while others do not. Facial comparison software of the same name is published as high risk by one of twelve bodies. No entry states why a class was chosen, so an external reader cannot distinguish defensible variation in deployment context from error. Separately, the schema provides no field able to hold an operational record, and in ten free-text fields no entry records an audit trail or a count of human corrections. We specify the conditions under which an instrument can be read as evidence of how the obligations that follow from a classification are carried out, not merely as a record of classified systems. The missing reasoning behind published classifications is the more tractable of the two gaps.

Keywords

EU AI Act, algorithm registers, AI governance, public-sector AI, accountability, deployer obligations, algorithmic transparency

1 Introduction

A public body preparing for the EU AI Act faces a sequence. The first step is to establish whether a system is high risk, because that determines which duties apply. The second is to discharge those duties, most of which concern the system in operation: assign oversight to persons with the necessary competence and authority, monitor use, retain the logs the system generates, and explain individual decisions on request. Since 2020, many bodies have also published an algorithm register, a public description of each system they operate [6, 7]. The instrument was widely praised [11].

The two are related in a way that has not been examined. The Dutch national register publishes a risk class for every entry, drawn from three values: high-risk AI system, impactful algorithm, and other. Only the first corresponds to a category in the Regulation; the other two are constructs of the domestic publication standard and carry no AI Act meaning. The register's own

front matter states that it covers impactful algorithms including high-risk AI systems, and the standard governing it records that no field is strictly mandatory because no legal obligation yet exists. The register is therefore where a body states publicly, in advance of any duty applying, whether it considers a system to fall in the Act's high-risk category.

We ask two questions. Is the classification reported by the register applied consistently between institutions and over time? And can an instrument of this class carry the operational duties that classification triggers?

Our finding is not that classification is wrong. Whether a system is high risk turns on intended purpose and context of use under Article 6 and Annex III, and two bodies running software of the same name in different contexts may correctly reach different answers. Our finding is that the register publishes the answer and never the reasoning. Divergence is therefore unreviewable: no external reader, including a supervisory authority, can tell a considered contextual judgement from an unconsidered one. The register records a confident single label over a question it does not show being asked.

We chose the strongest available case. The Dutch register descends from the Amsterdam model that many European registers follow, and nothing here should be read as criticism of either that city or the Dutch programme.

Our evidence is European, and the duties we decompose are the Act's, but the mechanism is not specific to either. Wherever a legal regime conditions an institution's duties on a risk class that the institution itself publishes, the same question arises: can anyone outside that institution review how the class was reached? Registers like this exist outside Europe, Canada's among them [8], and the first requirement we derive does not depend on the AI Act.

2 Background and Related Work

2.1 Registers and their critics

Floridi read the Amsterdam and Helsinki launches as evidence that AI could be governed as a public service [11]; Cath and Jansen replied that the registers were thin and selectively scoped [4]. Nieuwenhuizen found Dutch disclosures of limited use for external accountability while exerting a disciplining effect internally [16]. Das et al. audited Canada's federal register and showed its categories obscure the operational layer and the exercise of human discretion [8]. Norval et al. argue that disclosure supports accountability only when designed for the party who must act on it [17]. The Dutch data protection authority reported that more than half of Dutch municipalities had registered nothing and that fundamental rights assessments appeared in five percent of entries [3]. Two Dutch audits come closer to the register itself. The Court of Audit inventoried 433 AI systems at 70 central government organisations. For more than half, opportunities had not yet been weighed against risks. Most were rated minimal

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Information Society 2026, Ljubljana, Slovenia

© 2026 Copyright held by the owner/author(s).

<https://doi.org/10.70314/is.2026.ai4sdg.160>

risk. Only five percent appeared in the public register [1]. Algorithm Audit examined the inventories of fourteen ministries, 370 algorithms in all. It found about 3.5 percent classed as high-risk AI, against a Commission estimate of around 15 percent. It questioned whether the inventories had been made consistently [2]. Both examined what bodies reported about their own systems. Neither compared the class given to the same system by different bodies, or followed a class through the register's history.

This literature establishes that registers are incomplete and of limited accountability value. Earlier work has not mapped register content to the duties the Act puts on deploying bodies.

2.2 Records, evidence and explanation

A second literature concerns whether records support the findings asked of them. Janssen argues that a record satisfying the letter of a logging requirement can still fail to yield legally operative findings about oversight [13]. Chappidi et al. show empirically that record-keeping introduced for accountability reshapes the oversight it documents [5]. Neither literature has been applied to registers.

2.3 The duties and their timing

Article 26 sets twelve paragraphs of obligations for deployers of high-risk systems, including oversight by competent persons, monitoring and suspension on risk, retention of automatically generated logs where they are under the deployer's control, and notification of workers and affected persons [9]. Article 27 requires public bodies to conduct a fundamental rights impact assessment before first use and to notify the result to the market surveillance authority. Article 86 gives a person adversely affected by a decision based on a high-risk system's output the right to obtain clear and meaningful explanations of the system's role and the main elements of the decision. All are conditional on the classification. The Act addresses the classification itself, but at provider level. Under Article 6(4), a provider that considers an Annex III system not to be high-risk must document that assessment before placing the system on the market or putting it into service, and make it available to national authorities on request [9]. The reasoning sits with the provider and, on request, with the authority. For classification, the register shows only the class a body assigns.

The Digital Omnibus, Regulation (EU) 2026/1744, deferred the application of Annex III high-risk duties to 2 December 2027 [10]. It changed when the obligations apply rather than what they require. It also allows the assessment under Article 27 to draw on a data protection impact assessment. That changes how the duty is met, not whether it arises.

3 Method

3.1 Decomposition

We reduce the deployer duties of Articles 26 and 27, together with the Article 86 explanation right, to observable properties (Table 1), each classified as documentation-satisfiable or runtime-required. The step is analytical: it follows from reading the provisions, not from observing registers, and should be read as an organising instrument rather than a result. Article 26(10) is excluded as specific to law enforcement biometrics, and 26(3) is a saving clause that imposes no duty; all other paragraphs are represented.

3.2 Audit

We audit at three levels. At schema level we classify all 32 fields of publication standard v1.0 [15] by the content each can hold. At content level we compute completeness, classification distribution and cross-institutional consistency across all 1,534 entries. At revision level we analyse the complete published history of the register: 2,921 versions, 638 of the systems carrying more than one, giving 1,365 transitions. The history covers 1,556 distinct systems against 1,534 currently published entries; the difference is 22 systems present in the published history and since removed from the register, sixteen of them recorded as in use at their last publication.

Text comparison uses whitespace-normalised strings. We separate transitions from empty to populated fields, which reflect schema rollout, from rewrites of existing text. Cross-institutional comparison matches on entry name, lowercased and whitespace-trimmed, with no fuzzy matching. Exact matching under-counts, since the same product registered under a vendor variant or a differing local name is not detected, so the divergence figures below are a floor rather than an estimate.

The register is written in Dutch, and all content searches use Dutch terms with common inflections, including logboek, audit-trail, controlespoor, bewaartermijn, handmatige correctie and afwijking van het advies, supplemented by manual reading of matched entries in the original language with machine translation assistance, as declared below. All data were retrieved on 13 August 2026 from the register's own bulk export.

3.3 Scope

We do not assess whether any institution is compliant, nor treat a published category as a determination of legal status. Duties attach by operation of law under Article 6 and Annex III whatever a body publishes. Our claims concern the register: what it reports, how consistently, and what it can carry.

Institutions hold governance artefacts beyond the register: records of processing under the GDPR, impact assessment files, procurement contracts, internal audit trails. A register is not the only place accountability could live, and we do not claim otherwise. It is the only one that is public, and it is where the risk position is stated publicly.

4 Findings

4.1 Divergence is published without reasoning

Of 1,534 entries, 798 are published as other algorithms, 695 as impactful, and 41, some 2.7 percent, as high-risk AI systems. Of those 41, eight record any impact assessment. That is 19.5 percent, against a register-wide rate of 22.0 percent.

Among 111 system names published by more than one body, 43 carry more than one class, covering 451 entries. Seventeen of the 43 names have two publishers, so any disagreement is a tie. Among the 26 with three or more, the most common class covers a median of 78 percent of entries, and under 70 percent for nine (Figure 1).

Only one of the three classes carries AI Act meaning, so we separate the legally consequential subset. In eight of the 43 names, covering 91 entries, some bodies declare a system high risk while others do not. These include facial comparison software of the same name, published as high risk by one of twelve bodies, and a commercial assistant published as high risk by three of twelve.

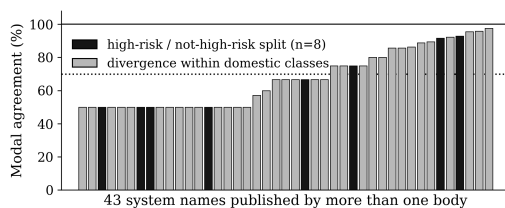


Figure 1: Modal agreement, meaning the share of entries in the most common class, for the 43 names, ascending. Two-publisher names sit at 50 percent by construction. The dotted line is at 70 percent. The dark bars are the eight names split on high-risk status.

Annex III lists remote biometric identification and excludes checking that a person is who they claim to be. Which of the two applies depends on context, and the register records no such context.

We do not claim the divergent entries are misclassified. Classification turns on intended purpose and context of use, and variation may be correct. We claim something narrower and, for a supervisory reader, more useful: the register states a class and never the reasoning behind it, so defensible variation and error are indistinguishable from outside. The instrument that would surface a misclassification is the one that records why the class was chosen, and it does not exist.

4.2 Reclassification is published without reasoning

The register documents 41 reclassifications between two explicitly stated categories. That number happens to match the earlier count of high-risk entries, but the two counts are about different observations. Twenty-seven move to a less stringent class, and ten of those twenty-seven leave the high-risk class. In 35 of the 41, neither the human oversight text nor the risk text changed.

That count requires its base rate, which we state plainly because it works against the reading the number invites. Across all 1,365 transitions, the three governance fields (human intervention, risks, and proportionality) are rewritten in 11.7 percent of cases. The four descriptive fields (short description, goal, source data, and methods) are rewritten in 13.7 percent. Under no association, roughly five of 41 reclassifications would carry a governance rewrite; six did. Reclassification is therefore not accompanied by less governance revision than an ordinary republication, and we make no claim that it is. With $n=41$ we can draw no inference about a difference in either direction.

What remains is descriptive and, we think, sufficient. Ten systems left the Act's high-risk category, including a generative assistant, an alimony calculator, automated debt remission and a document anonymisation tool. Nine of the ten record no revision to what they say about oversight or risk, and none records why the class changed.

4.3 The schema cannot hold an operational record

Publication standard v1.0 defines 32 fields: six identifiers, seven taxonomy or enumerated values, four URLs, three dates, two metadata fields and ten free-text descriptions. No field is structured to hold a log, an interaction, an oversight event, an incident, or the model version in force at a given use. This is a property of the standard, into which 336 organisations publish.

4.4 Description is populated, assessment is not

The goal field is completed in 99.5 percent of entries, human intervention in 98.4, and risks in 92.0. The field recording which impact assessments were performed reaches 22.0 percent; among the 41 self-declared high-risk systems, eight. A decommissioning date is recorded for 79 of the 1,534 entries, or 5.1 percent of the register. Of the 65 entries marked as out of use, 35 have one. The 22.0 figure is an upper bound: inspection of completed values finds content that is not an assessment, including one entry recording that the system speeds up the case handler's work.

Article 27 is the duty this measures against, and it is documentation-satisfiable. A register could carry it. That it does not is a matter of practice rather than of instrument design, and we treat it as the strongest case against our own argument.

4.5 Free text was available and was not used for operation

Ten free-text fields impose no restriction on content. Across all 1,534 entries, no entry mentions an audit trail and none records any count of human corrections or overrides of a system's output. Six state a log retention period, four state a monitoring frequency and two give an interaction count. Where institutions could have recorded operation in prose, they did not. Field definitions throughout are those of publication standard v1.0 [15].

4.6 A documented case

A system screening social assistance applications for investigation carries one of the most detailed entries in the register: no automated decision, three employees reviewing each case, work instructions written to counter automation bias, staff trained in interpreting output, and a fundamental rights assessment recorded. Its status changed to out of use in March 2025. Investigative reporting published in June 2025 documented discriminatory outcomes during its pilot [12]; the withdrawal preceded that publication and we make no causal claim. The oversight text was last edited in December 2024 and was not revised across three subsequent republications.

Every claim in that entry is a claim about practice, and the register carries no instance of any of them. The entry is not careless. It is careful, and careful description is what the instrument produces.

5 Derived Requirements

Two distinct gaps follow. The first concerns the classification itself, which is stated without evidence and revised without explanation. The second concerns the duties that classification triggers, which are discharged by events and cannot be evidenced by description. We state four requirements for an instrument that would address them. R1 and R4 follow from what we measured; R2 and R3 follow from the decomposition in Table 1 and from the statutory text. They are requirements, not a design.

R1. A stated risk class is accompanied by the reasoning and evidence that produced it, and any change to it records what changed and why. Divergence between institutions would then be visible and reviewable rather than silent. R1 asks the deploying body to publish what the Act asks only the provider to document, and only in private.

R2. Records are produced by the system as it operates rather than composed afterwards.

R3. Records are held under the institution's control. Article 26(6) conditions the retention duty on exactly this, and where

Table 1: Duty coverage by the register class. D: documentation-satisfiable; R: runtime-required; Reg.: whether the register carries it. Art. 26(10) excluded as law-enforcement specific.

Property	Art.	Layer	Reg.
Risk class stated	6 / III	D	yes
Risk class evidenced or reviewed	6 / III	D	no
Intended use and instructions documented	26(1)	D	yes
Use conformed to instructions	26(1)	R	no
Oversight arrangement described	26(2)	D	yes
Overseer competence evidenced	26(2)	D	no
Individual oversight act recorded	26(2)	R	no
Input data relevance verified per use	26(4)	R	no
Monitoring arrangement described	26(5)	D	yes
Monitoring performed, evidenced	26(5)	R	no
Suspension decision traceable	26(5)	R	no
Logs retained at least six months	26(6)	R	no
Workers informed, evidenced	26(7)	R	no
Registration in EU database	26(8)	D	no
DPIA informed by provider material	26(9)	D	partial
Affected person informed	26(11)	R	no
Cooperation with authority evidenced	26(12)	R	no
Impact assessment recorded (any)	27	D	≤22%
Individual decision explained on request	86	R	no

inference is served by a third party under contract, the condition is not obviously satisfied. The documentation for Microsoft Azure OpenAI, a widely used vendor platform, states that interaction data are retained in storage systems operated by the vendor outside the customer’s own tenant [14].

R4. Oversight is recorded as events rather than asserted as arrangements. The most informative available signal, how often a person rejected or amended a system’s output, is an event count, and no entry in the register reports one.

We stop here deliberately. Further requirements are readily proposed, on model version identity or record integrity for example, but nothing in this audit bears on them and we do not derive what we have not measured.

6 Limitations

Our audit covers one national register, among the largest in Europe and the ancestor of many. The UK, Canadian and French instruments were not examined.

The revision analysis covers the full published history but can only observe what the register records. We report the absence of published justification, not the absence of deliberation.

We audit published content, not institutional practice. An institution may oversee a system attentively and describe it poorly, or the reverse. Establishing the relationship between the two would require access we did not have.

7 Conclusion

The EU AI Act keys a public body’s duties to a classification, and most of those duties are discharged by events rather than description. The Dutch national register publishes a class for 1,534 systems without the reasoning behind it, while same-named systems carry different classes across institutions and ten systems left the high-risk class with no recorded reason. The same register has no field in which an event could be recorded, and where free text was available, no institution used it to record one.

The register does not evidence the risk position it reports, and it cannot evidence operation. An institution holding a complete and current register may still be unable to show which class a system belongs in, or to explain to a citizen why a decision went

against them. Whether an instrument meeting the requirements above can be built within the constraints public bodies operate under is the question we take up next.

Declarations

Data availability. All data are public. The analysis draws on the bulk exports of the Dutch national algorithm register retrieved 13 August 2026. The snapshot, derived tables and analysis code are deposited at <https://doi.org/10.5281/zenodo.22745470>, so that the figures reported here remain checkable as the live register changes.

Competing interests. The authors are affiliated with a company developing AI governance and infrastructure software. No product is named, evaluated or proposed in this paper. The requirements in Section 5 are derived from the audit reported here.

Generative AI use. Generative AI tools assisted with drafting, with the code supporting the data analysis, and with translation when reading Dutch-language register entries. All findings were computed from the cited public data, and all claims and references were verified by the authors, who take full responsibility for the content.

References

- [1] Algemene Rekenkamer. 2024. Focus op AI bij de rijksoverheid. (Oct. 16, 2024). Retrieved Sept. 28, 2026 from <https://www.rekenkamer.nl/publicaties/rapporten/2024/10/16/focus-op-ai-bij-de-rijksoverheid>.
- [2] Algorithm Audit. 2025. Inventory 14 Dutch ministries Netherlands algorithm registry. (Aug. 2025). Retrieved Sept. 28, 2026 from https://algorithmaudit.eu/knowledge-platform/knowledge-base/inventory_high_risk_ai_systems/.
- [3] Autoriteit Persoonsgegevens. 2025. Algorithm registration in the Netherlands needs improvement. (July 11, 2025). Retrieved Sept. 28, 2026 from <https://www.autoriteitpersoonsgegevens.nl/en/current/algorithm-registration-in-the-netherlands-needs-improvement>.
- [4] C. Cath and F. Jansen. 2022. Dutch comfort: the limits of AI governance through municipal registers. *Techné: Research in Philosophy and Technology*, 26, 3, 395–412.
- [5] S. Chappidi, J. Cobbe, C. Norval, A. Mazumder, and J. Singh. 2025. Accountability capture: how record-keeping to support AI transparency and accountability (re)shapes algorithmic oversight. In *Proc. AAAI/ACM AIES*. Vol. 8, 554–566. doi:10.1609/aies.v8i1.36570.
- [6] City of Amsterdam, City of Helsinki, and Saidot. 2020. Public AI registers: realising AI transparency and civic participation in government use of AI. White paper. (Sept. 2020).
- [7] City of Helsinki. 2026. AI register. Retrieved Aug. 12, 2026 from <https://ai.helsinki.fi/en/ai-register/>.
- [8] D. Das, C. Tessono, S. I. Ahmed, and S. Guha. 2026. Bureaucratic silences: what the Canadian AI register reveals, omits, and obscures. In *Proc. ACM FAccT*. doi:10.1145/3805689.3812336.
- [9] European Commission. 2026. AI Act Service Desk: articles 26, 27 and 86. Retrieved Aug. 13, 2026 from <https://ai-act-service-desk.ec.europa.eu>.
- [10] European Parliament and Council of the European Union. 2026. Regulation (EU) 2026/1744 amending Regulation (EU) 2024/1689. Official Journal of the European Union. (July 2026).
- [11] L. Floridi. 2020. Artificial intelligence as a public service: learning from Amsterdam and Helsinki. *Philosophy & Technology*, 33, 4, 541–546. doi:10.1007/s13347-020-00434-3.
- [12] E. Guo, G. Geiger, and J. Braun. 2025. Inside Amsterdam’s high-stakes experiment to create fair welfare AI. *MIT Technology Review*, (June 2025).
- [13] J. Janssen. 2026. From runtime records to legal findings: an evidentiary-adequacy criterion for agentic AI oversight. Preprint, not peer reviewed. (2026). arXiv: 2607.00941.
- [14] Microsoft. 2026. Data, privacy, and security for Azure OpenAI Service. Retrieved Aug. 13, 2026 from <https://learn.microsoft.com/legal/cognitive-services/openai/data-privacy>.
- [15] Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. 2026. Algoritmeregister van de Nederlandse overheid and publicatiestandaard 1.0. Retrieved Aug. 13, 2026 from <https://algoritmes.overheid.nl>.
- [16] E. Nieuwenhuizen. 2024. Algorithm registers: a box-ticking exercise or meaningful tool for transparency? *Information Polity*, 29, 4, 415–433. doi:10.1177/15701255241297107.
- [17] C. Norval, K. Cornelius, J. Cobbe, and J. Singh. 2022. Disclosure by design: designing information disclosures to support meaningful transparency and accountability. In *Proc. ACM FAccT*.

Sustainable Artificial Intelligence Infrastructure and Digital Sovereignty: A Proposal for a Sustainability Impact Assessment and a Data Center Responsibility Index

Leonardo David Quintiliano

Institute of Advanced Studies

University of São Paulo (USP) ESEG - ETAPA

São Paulo, SP, Brasil

leodaqui@usp.br

Abstract

The rapid expansion of generative artificial intelligence has increased demand for computing capacity, energy, water, connectivity, and storage and processing infrastructure. Brazil has relevant advantages for hosting new data centers, particularly its predominantly renewable electricity mix and strategic position in Latin America. Yet data center sustainability cannot be assessed solely through environmental and energy metrics, nor does the physical location of servers in Brazil by itself ensure digital sovereignty when technical or legal control remains subject to foreign jurisdictions. This article argues that sustainable AI infrastructure should integrate environmental sustainability, digital security and sovereignty, and economic and social sustainability. It proposes a Sustainability Impact Assessment (SIA) as an integrated system for licensing, public incentives, procurement, and the processing of government databases. Within the SIA, a Data Center Responsibility Index (DRI) is proposed as a quantitative metric based on six weighted indicators grouped into three axes. The combined SIA-DRI framework seeks to provide public authorities with an objective, auditable, and comparable instrument for authorizations, incentives, monitoring, and accountability.

Keywords

artificial intelligence, data centers, sustainability

1 Introduction

Artificial intelligence depends on extensive material infrastructure. Training and operating large-scale models require high-performance servers, graphics processing units, communication networks, storage systems, continuous power supply, and cooling mechanisms. The expansion of generative AI has therefore transformed data centers into strategic components of industrial, energy, and digital policy.

There is also a need to protect Brazilian personal data, particularly sensitive data, government databases, and strategic information assets. Infrastructure located within national territory can increase oversight capacity and reduce external dependencies, but it does not by itself solve the problem of digital sovereignty. A data center physically located in Brazil may remain under the technical or legal control of companies subject to foreign jurisdictions. The relevant question is therefore not

only where the server is located, but who controls the data, which laws may reach the provider, and where the information may be transferred or replicated.

Brazil's predominantly renewable electricity mix allows the country to present itself as a potential regional hub for low-carbon computing. Significant obstacles nevertheless remain, including long-distance connectivity, territorial concentration of facilities, dependence on imported equipment, shortages of specialized professionals, and the absence of an integrated regulatory framework [1].

Data center sustainability is commonly measured through energy efficiency, emissions, and water consumption. These factors are essential but insufficient. AI infrastructure processes large volumes of personal, strategic, and government data and produces significant economic and territorial effects. This article hypothesizes that its assessment must integrate environmental, legal, digital, economic, and social dimensions. To operationalize this approach, it proposes a Sustainability Impact Assessment (SIA) as a system for public analysis and decision-making. Within the SIA, the Data Center Responsibility Index (DRI) serves as a synthetic metric designed to make project outcomes comparable and auditable.

1.1 Method

The research is qualitative, exploratory, and documentary and uses the hypothetical-deductive method. The documentary corpus comprises four groups of sources: Brazilian rules on data protection and international data transfers; adequacy decisions between Brazil and the European Union; public documents concerning the expansion and environmental impacts of data centers in Brazil; and foreign rules allowing extraterritorial access to information controlled by technology companies. The analysis proceeded in three stages. First, identifying environmental, regulatory, and data-protection obligations. Second, grouping those obligations into three analytical axes. And third, converting the axes into six observable indicators. The SIA is the assessment system, combining mandatory legal requirements, gradable indicators, documentary transparency, and periodic review. The DRI is the quantitative metric used within that system.

2 The expansion of ai data centers in Brazil

Data centers are critical infrastructures in which data are stored, processed, and distributed. They support cloud computing, government applications, financial operations, communications, and AI systems. Brazil's National Telecommunications Agency (Anatel) recognizes their centrality to the economy and the continuity of digital services, requiring coordinated policies on security, resilience, and sustainability [2].

The computing density of servers equipped with graphics processing units increases demand for electricity and cooling. Some cooling technologies may also consume substantial quantities of water. In 2026, the Public Accounts Prosecutor's Office of the State of São Paulo filed a representation concerning

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Information Society 2026, 5–9 October 2026, Ljubljana, Slovenia

© 2025 Copyright held by the owner/author(s).

<http://doi.org/10.70314/is.2026.ai4sdg.155>

the environmental licensing of data center complexes in Sumaré and Hortolândia. The document questioned evaporative cooling systems and referred to estimates of water consumption of up to 3.24 million liters per day [3]. This was an allegation submitted to oversight bodies rather than a definitive conclusion, but the case illustrates local conflicts involving digital infrastructure, water, electricity, and environmental licensing.

Brazil's policy for encouraging the sector is still developing. Tax measures and legislative proposals designed to attract data centers seek to convert the country's energy advantage into computing capacity [4]. Public incentives, however, should not depend only on the amount invested or generic claims regarding renewable energy. Tax benefits should be conditioned on efficient use of natural resources, verifiable technological and social benefits, and safeguards concerning security, governance, and data control.

3 Data location, digital sovereignty, and foreign jurisdictions

Installing servers in Brazilian territory may reduce latency, improve service continuity, facilitate oversight, create skilled jobs, and expand the computing capacity available in the country. The physical location of equipment, however, is not the same as legal control over data.

A data center located in Brazil may belong to a company headquartered abroad or provide services to it. That company may control access keys, operating systems, contracts, and the global storage architecture. Data may also be replicated, accessed remotely, or retransferred to other facilities. A previous study on international transfers of Brazilian personal data by Big Tech companies showed that transnational data circulation challenges State sovereignty and informational self-determination, particularly when data sent to one jurisdiction are later transferred to companies or countries providing different levels of protection [5].

The U.S. CLOUD Act illustrates the limits of a policy based only on location. Providers subject to United States jurisdiction may be required to preserve or disclose data within their possession, custody, or control regardless of the country in which the information is stored [6]. Thus, servers physically located in Brazil or the European Union may remain exposed to foreign orders depending on the provider's corporate structure and degree of control over the information.

This does not mean that foreign providers should be prohibited or that every governmental request should be presumed illegitimate. It means that regulatory assessment must identify jurisdictional conflicts, procedures for responding to public-authority orders, and safeguards available to data subjects. Digital sovereignty should be understood as the State's effective capacity to establish rules, supervise operations, hold actors accountable, and protect relevant data. Territorial location without transparency concerning control and access may create only an appearance of sovereignty.

4 Data transfers and onward transfers after Brazil-EU adequacy

Resolution CD/ANPD No. 19/2024 regulated the international transfers provided for in Brazil's General Data Protection Law and established mechanisms including adequacy decisions, standard contractual clauses, specific contractual clauses, and binding corporate rules [7]. In 2026, Brazil and the European Union adopted reciprocal adequacy decisions, reducing barriers to data flows and strengthening convergence between their protection regimes [8], [9].

Adequacy reduces transaction costs and may facilitate the international activity of Brazilian companies while reinforcing the institutional credibility of the Brazilian Data Protection Authority (ANPD) [10]. The regulatory problem, however, does not end with identifying the first destination. Brazilian data lawfully transferred to the European Union may subsequently be processed by companies subject to third-country jurisdictions. The question becomes how to prevent circulation between systems considered adequate from operating as a general authorization for uncontrolled onward transfers.

European experience confirms this concern. The EU-U.S. Data Privacy Framework contains specific requirements for onward transfers [11]. Schrems I and Schrems II demonstrate the instability of international arrangements when surveillance rules and available remedies fail to provide essentially equivalent protection [12], [13]. The subsequent adoption of Executive Order 14086 and a new transatlantic mechanism restored greater legal certainty without removing the need for continuous review [11], [14].

In contemporary cloud architectures, onward transfer may be structural. Training data, prompts, logs, backups, security telemetry, and model outputs may be processed by multiple entities and replicated across several regions. Operators and providers should therefore disclose storage and replication locations, subprocessors, countries involved, potentially applicable foreign laws, and governmental access requests.

5 Data centers are necessary, but they need limits - the Sustainability Impact Assessment (SIA) and the Data Center Responsibility Index (DRI)

Data center deployment is both an economic strategy and a sovereignty issue. The proposed Sustainability Impact Assessment (SIA) integrates environmental, data-protection, cybersecurity, and socioeconomic review for large facilities. Its quantitative component, the Data Center Responsibility Index (DRI), converts auditable indicators into a 0-100 score while preserving disqualifying legal and security requirements.

The DRI has three axes: (1) environmental sustainability, measured through energy efficiency, renewable-energy share, carbon intensity, and local environmental pressure [15]-[18]; (2) digital security and sovereignty, covering access control, encryption, incident response, storage and replication transparency, subprocessors, jurisdictions, and foreign-law exposure [7], [19]; and (3) economic and social sustainability, measured through qualified jobs, local procurement, and computing capacity available in Brazil. Tables 1 and 2 reproduce the proposed pilot scales.

5.1 Axis 1 - environmental sustainability

Axis 1 evaluates environmental performance through four indicators: energy efficiency (PUE and kWh/TB), renewable-energy factor (REF), carbon intensity (CUE), and local environmental pressure, including water use (WUE) adjusted to local water stress. Combining these measures prevents strong performance in one environmental dimension from masking weaknesses in another [15]-[18].

The environmental axis deliberately combines relative efficiency with absolute and territorial pressure. PUE identifies the energy consumed by supporting infrastructure, but it does not measure computing productivity or the carbon profile of the electricity used. REF therefore captures the renewable share, while CUE connects consumption to emissions. WUE and the Local Environmental Pressure Index complement these indicators by incorporating water stress, effects on ecosystems and land,

residual heat, refrigerants, waste, and other externalities. This contextualization is important because the same technical efficiency may represent very different environmental risks in water-abundant and water-stressed regions [20]-[23].

No single environmental metric captures the full sustainability profile of a data center. PUE identifies facility overhead, but not computing productivity, the carbon profile of electricity, or pressure on local water resources. REF and CUE therefore complement PUE by identifying the renewable origin and climate effects of electricity, while WUE adds the water dimension. The proposed Local Environmental Pressure Index also considers effects on ecosystems and land, residual heat, waste, refrigerants, and other externalities. Activity-normalized measures should be compared only among technically similar facilities, because one terabyte used for storage is not computationally equivalent to one terabyte processed in intensive AI training [20]-[23].

Table 1: Environmental sustainability indicators and pilot scale

Indicator	Metric	Pilot scale
Energy	PUE + kWh/TB, compare kWh/TB with the category median.	PUE: $\leq 1.20=5$; $1.21-1.30=4$; $1.31-1.40=3$; $1.41-1.50=2$; $1.51-1.70=1$; $>1.70=0$.
Renewable energy	REF = renewable energy / total energy [16].	100%=5; 90-99.99%=4; 75-89.99%=3; 50-74.99%=2; 25-49.99%=1; $<25\%=0$.
Carbon	kgCO ₂ e/TB and CUE [17]; ratio to category median.	$\leq 50\%=5$; $50-70\%=4$; $70-90\%=3$; $90-110\%=2$; $110-130\%=1$; $>130\%=0$.
Local impact	LEPI= $0.40W+0.25E+0.20T+0.15R$; W uses WUE [18].	W, E, T, and R score 0-5; WUE is adjusted for local water stress.

5.2 Axis 2 - digital security and sovereignty

Axis 2 jointly evaluates technical security and legal control over data. It considers access control, encryption and key management, supported critical components, continuous monitoring and incident response, and transparency about storage and replication locations, subprocessors, countries involved, and potentially applicable foreign laws [7], [19].

For digital security and sovereignty, the model separates infrastructure protection from the broader question of legal control. The proposed indicator combines identity and access management, multifactor authentication, least privilege, encryption in transit and at rest, auditable key management, supported components, continuous monitoring, event logging, and tested incident response. It also requires identification of storage and replication locations, subprocessors, the countries involved, and potentially applicable foreign laws. This final component is essential because a facility located in Brazil may remain subject to external legal orders through the corporate structure of the provider. High environmental performance should therefore never conceal weak protection or uncertain legal control over strategic data [7], [19].

The security and sovereignty axis is designed as a cumulative set of controls rather than a single cybersecurity score. It requires access control, multifactor authentication, least privilege, encryption in transit and at rest, auditable key management, supported critical components, continuous monitoring, event logging, and a tested incident-response plan. These controls are combined with jurisdictional transparency: operators should identify storage and replication locations, subprocessors, countries involved, and potentially applicable foreign laws. This reflects the distinction between territorial location and effective legal control. The objective is not to exclude foreign providers, but to make conflicts of jurisdiction visible before strategic or governmental data are entrusted to the infrastructure [7], [19].

5.3 Axis 3 - economic and social sustainability

Axis 3 measures lasting economic and social returns through qualified employment (QES), local procurement (LPS), and domestic computing capacity (DCCS). It asks whether the

project creates durable local linkages and makes computing capacity effectively available to Brazilian firms, universities, research centers, and public bodies.

The economic and social axis seeks to distinguish durable domestic benefits from the announced value of an investment. QES focuses on qualified direct jobs actually maintained during operation; LPS measures local purchases by economic value rather than by the number of contracts; and DCCS verifies the share of installed computing capacity effectively available to Brazilian companies, universities, research centers, and public bodies. The purpose is to connect public incentives to measurable local and technological benefits rather than to the physical presence of a facility alone.

The economic and social axis converts broad policy promises into measurable variables. QES compares qualified direct jobs with the state stock of qualified formal employment; LPS measures the annual economic value purchased from local suppliers rather than merely counting contracts; and DCCS measures the share of installed computing capacity effectively available to Brazilian companies, universities, research centers, or public bodies. Together, these measures seek to identify whether the project leaves a durable technological and productive footprint. Contributions to the Sustainable Development Goals should therefore count only when tied to verifiable targets, data sources, and outcomes rather than generic corporate commitments.

Table 2: Economic and social sustainability scoring

Metric	Score 5	Score 4	Score 3	Score 2	Score 1	Score 0
QES	$\geq 1\%$	0.50-0.99	0.25-0.49	0.10-0.24	0.01-0.09	$<0.01\%$
LPS	$\geq 60\%$	45-59%	30-44%	15-29%	5-14%	$<5\%$
DCCS	$\geq 50\%$	35-49%	20-34%	10-19%	$>0-9\%$	0%

Indicators receive scores from 0 to 5 and are weighted to produce the DRI. Table 3 illustrates the calculation; Table 4 presents the proposed classification. Missing licenses, unidentified controllers or jurisdictions with data access, unresolved critical vulnerabilities, unauditable data, or incompatible onward transfers of government or strategic data remain disqualifying regardless of the aggregate score [7].

Table 3: Illustrative DRI calculation

Indicator	Ni	Weight	Contribution
Efficiency	4	20	16
Renewable energy	5	15	15
Carbon	3	10	6
Local impact	2	15	6
Security/sovereignty	5	25	25
Social impact	3	15	9
DRI	—	100	77

Table 4: DRI classification and suggested effect

DRI	Classification	Suggested effect
90-100	Exceptional	Priority
80-89	High	Fully eligible
70-79	Adequate	Eligible
60-69	Intermediate	Improvement plan
<60	Insufficient	No incentive

5.4 Calculation, disqualifying criteria, and classification

The illustrative calculation in Table 3 shows how the weighted score is produced, but the aggregate result must be read together with the three axes. A strong score in one dimension may otherwise conceal a critical weakness in another. For that reason, the model does not permit unlimited trade-offs. Missing licenses, unidentified controllers or jurisdictions, unsupported critical components, unaddressed critical vulnerabilities, unauditable data, or incompatible onward transfers of government or strategic data prevent favorable treatment regardless of the numerical score. Table 4 therefore functions only after these threshold conditions have been satisfied.

The numerical result should be treated as an aid to decision-making, not as an automatic authorization. In the illustrative case, scores of 4 for efficiency, 5 for renewable energy, 3 for carbon, 2 for local impact, 5 for security and sovereignty, and 3 for economic and social impact produce a DRI of 77. The overall result remains subject to examination of the individual axes, since a serious local environmental weakness should not disappear behind stronger performance elsewhere.

5.5 Application, review, and validation

Administrative application of the SIA may follow four steps: submission of primary data and measurement methods; verification of mandatory information and disqualifying criteria; calculation of the component scores and the DRI; and publication of a non-confidential result with the sources and assumptions used. For projects receiving public incentives or processing government data, review should be periodic and should also occur after material changes in energy mix, cooling technology, corporate structure, subprocessors, storage countries, or security incidents. The SIA thereby operates both *ex ante* and *ex post*.

Periodic review also reduces regulatory gaming. Operators should not be able to select only their most efficient period, exclude auxiliary loads, or alter denominators to improve kWh/TB artificially. Measurement boundaries, the reference period, and the technical comparison category should therefore be defined in advance. When methodology changes, previous results should be recalculated whenever feasible, preserving a historical series that allows authorities to compare promised performance at authorization with verified performance during operation and tie incentives accordingly.

The assessment should also be dynamic. For projects receiving tax incentives, public contracts, or access to government databases, the DRI should be updated periodically and reviewed after material changes in energy mix, cooling technology, corporate structure, subprocessors, storage countries, security incidents, or domestic computing capacity. Comparing promised performance with verified operation allows incentives to be tied to actual results. The initial weights and ranges remain normative and should be tested on heterogeneous facilities through inter-rater reproducibility, sensitivity analysis, correlation with observed impacts, and recalibration by technical category and climate region [20], [21]. Such validation is essential before the index is treated as a stable regulatory benchmark rather than as a transparent prototype for policy experimentation.

The Sumaré and Hortolândia cases illustrate the model: the reported water-consumption estimate warrants scrutiny, but the absence of public data on the other indicators prevents a reliable score [3]. This information deficit is itself relevant for regulatory decision-making. The DRI therefore remains a prototype requiring empirical validation and recalibration [20], [21].

6 Governance, incentives, and the sustainable development goals

Governance of the SIA could involve coordinated action by the ANPD, Anatel, environmental, energy-policy, competition, and information-security authorities. A public, non-confidential DRI would support oversight and comparison between promised and actual impacts. Public incentives should be tied to verifiable targets, with suspension, corrective obligations, or repayment in cases of subsequent non-compliance.

The framework also links digital infrastructure to SDGs 6, 7, 9, 12, 13, and 16 through water and energy efficiency, resilient technological capacity, responsible resource use, climate performance, and accountable data governance.

7 Conclusion

Brazil has relevant conditions to become a hub for low-carbon computing infrastructure, but sustainability cannot be reduced to investment or server location. Renewable energy may coexist with excessive local water use, technological dependence, or exposure of Brazilian data to foreign legal regimes. AI infrastructure must therefore be assessed as a multidimensional regulatory problem.

Brazil-EU adequacy is an important advance, but onward transfers and jurisdictional conflicts still require monitoring. The SIA integrates these concerns, while the DRI converts environmental, security, sovereignty, and socioeconomic factors into auditable indicators. Its principal contribution is to make explicit which information the State should require before granting incentives or treating infrastructure as sustainable.

The combined SIA-DRI framework is intended not merely to rank projects, but to make the conditions for public support explicit and auditable. In practice, sustainability requires efficient resource use, resilience, data protection, transparency over international data flows, regulatory capacity, and verifiable domestic technological and social benefits.

References

- [1] M. E. F. Viana and O. F. C. R. Scheidegger, "Infraestrutura computacional para a era da IA: desafios, políticas e oportunidades em data centers no Brasil" [Computing infrastructure for the AI era: challenges, policies, and opportunities for data centers in Brazil], BNDES Setorial, vol. 31, no. 60, pp. 167-205, 2025.
- [2] Brazilian National Telecommunications Agency (Anatel), White Paper Data Centers, Brasília, Brazil: Anatel, 2025.
- [3] Public Accounts Prosecutor's Office of the State of São Paulo, "Representação para apuração de possíveis irregularidades no licenciamento ambiental de data centers em Sumaré e Hortolândia" [Representation concerning possible irregularities in the environmental licensing of data centers in Sumaré and Hortolândia], São Paulo, Brazil, 2026.
- [4] Brazil, Chamber of Deputies, Bill No. 278 of 2026, establishing the Special Tax Regime for Data Center Services (REDATA).
- [5] L. D. Quintiliano and J. C. da Silva, "Limitações jurídicas à transferência internacional de dados pessoais brasileiros pelas Big Techs" [Legal limitations on international transfers of Brazilian personal data by Big Tech companies], Revista Jurídica da Presidência, vol. 28, no. 144, pp. 132-166, 2026, doi: 10.20499/2236-3645.RJP2026v28e144-3328.
- [6] U.S. Department of Justice, The Purpose and Impact of the CLOUD Act. Washington, DC, USA, 2019.
- [7] Brazilian National Data Protection Authority (ANPD), Resolution CD/ANPD No. 19 of Aug. 23, 2024, Regulation on International Data Transfers.
- [8] Brazilian National Data Protection Authority (ANPD), Resolution CD/ANPD No. 32 of Jan. 26, 2026, Adequacy Decision Concerning the European Union.
- [9] European Commission, Commission Implementing Decision pursuant to Regulation (EU) 2016/679 on the adequate protection of personal data by Brazil, C(2026) 373 final, Jan. 26, 2026.
- [10] L. D. Quintiliano, "Transferência internacional de dados entre Brasil e UE: o dilema da ANPD após a decisão de adequação" [International data transfers between Brazil and the EU: the ANPD's dilemma following the adequacy decision], Migalhas, Sep. 12, 2025.
- [11] European Commission, Commission Implementing Decision (EU) 2023/1795 on the adequate level of protection of personal data under the EU-U.S. Data Privacy Framework, Jul. 10, 2023.
- [12] Court of Justice of the European Union, Case C-362/14, Maximilian Schrems v. Data Protection Commissioner, Judgment of Oct. 6, 2015.
- [13] Court of Justice of the European Union, Case C-311/18, Data Protection Commissioner v. Facebook Ireland Ltd. and Maximilian Schrems, Judgment of Jul. 16, 2020.
- [14] United States, Executive Order 14086, Enhancing Safeguards for United States Signals Intelligence Activities, Oct. 7, 2022.
- [15] ISO/IEC 30134-2, Information technology — Data centres key performance indicators — Part 2: Power usage effectiveness (PUE), 2026.
- [16] ISO/IEC 30134-3, Information technology — Data centres key performance indicators — Part 3: Renewable energy factor (REF).
- [17] ISO/IEC 30134-8, Information technology — Data centres key performance indicators — Part 8: Carbon usage effectiveness (CUE).
- [18] ISO/IEC 30134-9, Information technology — Data centres key performance indicators — Part 9: Water usage effectiveness (WUE).
- [19] National Institute of Standards and Technology, NIST SP 800-207, Zero Trust Architecture, 2020.
- [20] X. Shao, Z. Zhang, P. Song, Y. Feng, and X. Wang, "A review of energy efficiency evaluation metrics for data centers," Energy and Buildings, vol. 271, art. 112308, 2022, doi: 10.1016/j.enbuild.2022.112308.
- [21] N. Homer and I. Azevedo, "Power usage effectiveness in data centers: overloaded and underachieving," The Electricity Journal, vol. 29, no. 4, pp. 61-69, 2016, doi: 10.1016/j.tej.2016.04.011.
- [22] D. Mytton, "Data centre water consumption," npj Clean Water, vol. 4, art. 11, 2021, doi: 10.1038/s41545-021-00101-w.
- [23] E. Masanet, A. Shehabi, N. Lei, S. Smith, and J. Koomey, "Recalibrating global data center energy-use estimates," Science, vol. 367, no. 6481, pp. 984-986, 2020, doi: 10.1126/science.aba3758.

Equitable Urban Flood Risk Mapping with Ensemble Machine Learning and Google Earth Engine: A Responsible AI Approach for Climate-Resilient Cities in Data-Scarce Regions

Peter Damilare Idowu[†]

Department of Urban and Regional Planning, Osun State Polytechnic
Iree, Osun State, Nigeria

M.Tech. Candidate, Urban and Regional Planning (Disaster and Environmental Management option)
Federal University of Technology, Akure, Ondo State, Nigeria
damilarepitaz@gmail.com

Abstract

Climate-driven urban flooding disproportionately affects informal settlements and low-income communities in the Global South, undermining Sustainable Development Goals (SDGs) 11 (Sustainable Cities and Communities) and 13 (Climate Action). Existing flood risk models often depend on high-resolution commercial data or dense gauge networks that are unavailable in many rapidly urbanizing regions. This paper presents an open-source, reproducible pipeline that combines multi-temporal Google Earth Engine (GEE) predictors with XGBoost and Random Forest classifiers to generate equity-aware flood susceptibility maps for a secondary city in southwestern Nigeria. Applied to the Ala, Owuru, and Elegbaka river corridors in Akure, the models achieved AUC-ROC values of 0.92 (XGBoost) and 0.89 (Random Forest) on spatially held-out data. A post-processing equity layer combining population density, informal settlement indicators, and access to early-warning infrastructure identifies neighborhoods where physical hazard and social vulnerability coincide but which hazard-only maps overlook. The workflow uses only free data and is designed for transfer to other data-scarce cities, offering municipalities a low-cost pathway to justice-oriented flood-risk intelligence.

Keywords

flood risk mapping, ensemble machine learning, Google Earth Engine, climate resilience, equitable AI, SDG 11, SDG 13, disaster risk reduction, responsible AI

1 Introduction

Urban flooding is among the most pervasive and rapidly intensifying climate hazards facing secondary cities across sub-

Saharan Africa. Rapid, largely unplanned urbanization, degradation of natural drainage corridors, and increasingly erratic rainfall place informal settlements and low-income neighborhoods at disproportionate risk [8]. In Nigeria, cities such as Akure exemplify this challenge: encroachment along river corridors, inadequate drainage, and weak enforcement of setback regulations have turned seasonal rainfall into recurring disasters. These dynamics directly undermine SDG 11, which calls for inclusive, safe, resilient, and sustainable human settlements, and SDG 13, which calls for urgent action on climate change and its impacts.

Conventional flood risk assessment relies on hydrological models calibrated against dense rain and stream gauge networks, or on high-resolution commercial imagery. Both are scarce or absent in many rapidly urbanizing regions of the Global South, leaving municipalities without the evidence base for proactive disaster risk reduction. Machine learning trained on free Earth observation data is a promising alternative, but existing applications rarely include an explicit equity lens that separates areas of high physical hazard from areas where hazard intersects with social vulnerability.

This paper addresses that gap with an ensemble machine learning pipeline built entirely on free and open data within Google Earth Engine (GEE), producing both flood susceptibility scores and an equity-adjusted risk surface. The contribution is threefold: (i) a reproducible, low-cost workflow transferable to other data-scarce cities; (ii) an empirical demonstration on the Ala, Owuru, and Elegbaka river corridors in Akure, Nigeria; and (iii) an equity-adjustment step that foregrounds distributional justice in how flood risk intelligence is generated and communicated, in line with responsible, human-centered AI for the public good.

2 Related Work

Machine learning is now widely used for flood susceptibility mapping, from ensemble models to deep learning frameworks [1, 2]. Tree-based methods, notably Random Forest [7] and gradient boosting [6], are popular because they capture nonlinear interactions among topographic, hydrological, and land-cover predictors. Cloud platforms such as GEE lower the barrier for researchers in low-resource settings by providing free access to petabyte-scale satellite archives [3], including

[†]Corresponding author.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, 5–9 October 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
<https://doi.org/10.70314/is.2026.ai4sdg.180>

Sentinel-1 synthetic aperture radar (SAR), which is widely used to delineate historical flood extents even under cloud cover [5].

Two gaps remain. First, much of this literature is developed and validated where in-situ monitoring is dense, while African secondary cities, whose growth is rapid and poorly monitored [8], receive comparatively little attention. Second, most susceptibility maps report only physical hazard probabilities and do not link them to the exposure and adaptive capacity of the people living in high-risk zones [9]. Scholarship on responsible and justice-oriented AI argues that hazard intelligence divorced from vulnerability context risks reproducing existing inequities in how disaster resources are allocated [9]. Susceptibility mapping itself is well established; the contribution here lies in combining it with an explicit, open, and reproducible equity layer for a data-scarce African city, and in validating the models with spatially disjoint folds.

3 Study Area and Data

3.1 Study Area

The study covers Akure, the capital of Ondo State, Nigeria, and specifically the catchments of the Ala, Owuru, and Elegbaka rivers, which traverse the urban core and have been the sites of repeated flood events in recent years. Rapid population growth, encroachment onto floodplains and wetlands, and inadequate stormwater infrastructure make Akure representative of rapidly urbanizing, data-scarce secondary cities in the region [8].

3.2 Predictors and Labels

The feature set was assembled entirely from free and open data processed in GEE at 30 m spatial resolution [3] (Table 1). Topographic predictors (elevation, slope, and the topographic wetness index) were derived from digital elevation data. Land-cover dynamics came from multi-temporal classification, and precipitation statistics from the CHIRPS dataset [4]. Proximity to mapped drainage networks captured channel-adjacency effects. Historical flood extents used as training labels were derived from Sentinel-1 SAR change detection [5] and supplemented with community-reported flood events, with stratified spatial sampling to balance flooded and non-flooded classes.

Table 1: Inputs used in the pipeline (all free and open).

Group	Variables	Role / source
Terrain	Elevation, slope, topographic wetness index	Predictor; digital elevation data in GEE
Rainfall	Precipitation statistics, cumulative anomalies	Predictor; CHIRPS [4]
Land cover	Multi-temporal land-cover dynamics	Predictor; GEE classification
Hydrology	Proximity to drainage network	Predictor
Flood labels	Historical flood extents	Sentinel-1 SAR [5] + community reports
Social vulnerability	Population density, informal settlements, early-warning access	Equity layer (post-processing) [9]

4 Methodology

4.1 Susceptibility Modelling

XGBoost [6] and Random Forest [7] classifiers were trained on the feature set above. Model development used stratified k-fold cross-validation with spatial blocking, in which training and validation folds are spatially disjoint, to prevent spatial autocorrelation from inflating apparent skill. Hyperparameters were tuned by grid search on the training folds, optimizing area under the receiver operating characteristic curve (AUC-ROC).

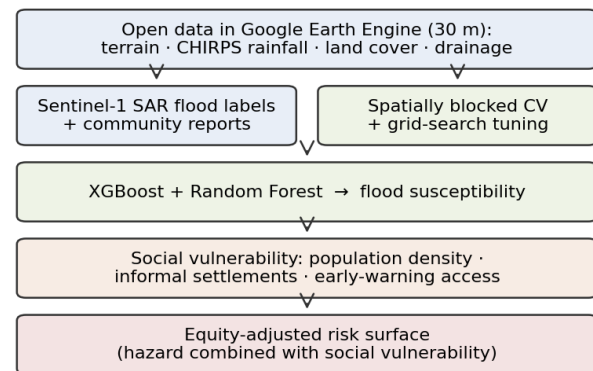


Figure 1: Workflow from open data to the equity-adjusted risk surface.

4.2 Equity Adjustment

The predicted hazard surface was then combined with three social vulnerability indicators drawn from established frameworks [9]: population density; an informal settlement indicator derived from built-form and land-use classification; and a proxy for early-warning access based on proximity to communication and emergency-response assets. The resulting composite, the *equity-adjusted risk surface*, highlights locations where elevated physical hazard coincides with elevated social exposure and separates them from locations of high hazard but low vulnerability. The whole pipeline (Figure 1) was containerized so that local technical teams with modest computing resources can run it.

5 Results and Discussion

5.1 Predictive Performance

On spatially held-out validation data, XGBoost reached an AUC-ROC of 0.92 and Random Forest 0.89 (Table 2), indicating strong discrimination of flood-prone areas and in line with the gains reported for ensemble methods in comparable settings [1]. Feature importance analysis showed that dynamic land-cover change and cumulative precipitation anomalies were consistently among the strongest predictors, ahead of static topographic variables alone. This underlines the value of multi-temporal Earth observation over one-time, terrain-based assessments.

Table 2: Discrimination on spatially held-out validation data.

Model	AUC-ROC	Validation
XGBoost	0.92	Spatially blocked, held-out
Random Forest	0.89	Spatially blocked, held-out

5.2 Equity-Adjusted Risk

The equity-adjusted surface flagged a number of neighborhoods along the Owuru and Elegbaka corridors that ranked only moderately on physical hazard alone but rose sharply in composite risk once population density and informal settlement extent were included. Conventional hazard-only maps tend to overlook exactly these areas, which post-flood assessments have repeatedly identified as sites of disproportionate loss, echoing findings that hazard and social vulnerability often diverge spatially [9]. This divergence between hazard-only and equity-adjusted rankings is the central empirical finding: physical and social risk are not interchangeable, and prioritization exercises that rely on the former alone risk misallocating scarce adaptation resources [8].

5.3 Limitations

Flood labels derived from SAR change detection and community reports are necessarily incomplete, and the socioeconomic indicators in the equity layer are proxies rather than direct measurements of vulnerability or adaptive capacity. Performance reflects Akure's hydrology and urban form and should be locally recalibrated before transfer to other cities. Even so, the consistency of the feature importance patterns with the flood risk literature, and the plausibility of the equity-adjusted results against known flood-affected neighborhoods, support the overall approach.

6 Conclusion and Future Work

This paper presented a reproducible, open-source pipeline that couples machine learning with GEE to produce flood

susceptibility and equity-adjusted risk maps for data-scarce secondary cities. In Akure, it achieved strong predictive performance while surfacing socially vulnerable neighborhoods that hazard-only mapping overlooks, giving municipalities and civil-society organizations a low-cost route to inclusive, justice-oriented adaptation planning.

Future work will (i) explore federated learning across several secondary cities to improve generalizability without centralized data sharing; (ii) integrate community-sourced validation to strengthen local ownership, ground-truth accuracy, and trust; and (iii) embed the equity-adjusted surface in municipal early-warning and resource-allocation workflows together with local disaster management authorities.

Acknowledgments

The author acknowledges the guidance of the M.Tech. supervisory committee at the Federal University of Technology, Akure, and the support of the Texas A&M University LAUP Pathways Mentorship Program.

References

- [1] M. S. Tehrany, B. Pradhan, and M. N. Jebur. 2014. Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS. *J. Hydrol.* 512 (2014), 332–343.
- [2] Y. Wang, Z. Fang, H. Hong, and L. Peng. 2020. Flood susceptibility mapping using convolutional neural network frameworks. *J. Hydrol.* 582 (2020), 124482.
- [3] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* 202 (2017), 18–27.
- [4] C. Funk, P. Peterson, M. Landsfeld, et al. 2015. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Sci. Data* 2 (2015), 150066.
- [5] A. Twele, W. Cao, S. Plank, and S. Martinis. 2016. Sentinel-1-based flood mapping: a fully automated processing chain. *Int. J. Remote Sens.* 37, 13 (2016), 2990–3004.
- [6] T. Chen and C. Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 785–794.
- [7] L. Breiman. 2001. Random forests. *Mach. Learn.* 45, 1 (2001), 5–32.
- [8] UN-Habitat. 2022. *World Cities Report 2022: Envisaging the Future of Cities*. United Nations Human Settlements Programme, Nairobi.
- [9] S. Rufat, E. Tate, C. G. Burton, and A. S. Maroof. 2015. Social vulnerability to floods: Review of case studies and implications for measurement. *Int. J. Disaster Risk Reduct.* 14 (2015), 470–486.

Generative AI Readiness for Future Skills and Lifelong Learning: Engineering and Non-Engineering Students at UTB and Universiti Malaya

Sharina Yunus[†]
Universiti Teknologi Brunei
Bandar Seri Begawan, Brunei
Darussalam
sharina.yunus@utb.edu.bn

Muhammad Sohail
Vilniaus Kolegija Higher
Education Institution
Vilnius, Lithuania
m.sohail@vuf.viko.lt

Nur Lisa Zaharin
Universiti Malaya
Kuala Lumpur, Malaysia
lisa.zaharin@um.edu.my

Abstract

Generative AI (GenAI) has become a routine part of university study, but frequent use does not necessarily mean that students use it critically. This exploratory cross-sectional study surveyed 118 students from two contrasting cohorts: 86 engineering students at Universiti Teknologi Brunei (UTB) and 32 non-engineering students at Universiti Malaya (UM). Composite measures (Cronbach's $\alpha = .74-.96$) covered frequency and breadth of use, perceived competence, metacognitive self-assessment and verification behaviour. For RQ1, Mann-Whitney U tests found no significant cohort differences except in verification behaviour, where UM students scored higher ($p = .010$, $r = .33$); this difference did not survive correction for six comparisons. For RQ2, frequency of use was strongly associated with breadth of academic use ($p = .779$, $p < .001$) but only weakly with verification behaviour ($p = .205$, $p = .037$, uncorrected) and not with metacognitive self-assessment ($p = .143$, $p = .139$). Perceived competence, metacognition and verification were moderately associated with one another ($p = .413-.502$, $p < .001$). How often students use GenAI therefore says little about how critically they use it. Higher education institutions need to go beyond rules on when AI may be used and give students structured practice in verifying AI-generated information and forming their own judgements.

Keywords

AI literacy, generative AI, higher education, learner autonomy, lifelong learning, responsible AI use.

1 Introduction

The use of generative AI (GenAI) has become widespread among university students. With answers available almost

instantly, students may now consult GenAI before turning to a textbook, lecture notes or other academic sources. This ease of access has changed how students search for information and approach their academic work. However, easy access to an answer does not necessarily mean that students are able to judge whether the answer is accurate, reliable or appropriate for the task. At the same time, students and educators have raised concerns about accuracy, ethics and over-reliance [1, 2]. Research has also shown that over-reliance on AI-generated information may affect important skills such as critical thinking, decision-making and independent judgement [3].

The important question for higher education is therefore not only how often students use GenAI, but how they use it. Frequent use may show familiarity with the technology, but it does not necessarily show that students are checking the information produced, questioning weak answers or deciding when GenAI is appropriate. Recent research on AI-supported self-regulated learning has similarly highlighted the importance of metacognition, ethical awareness and learner autonomy when students use AI for learning [4], [5]. UNESCO also emphasises a human-centred approach to AI in education, where students are expected to use AI responsibly and develop critical judgement rather than simply depend on the technology [6, 7]. OECD guidance for higher education has also identified the reliability of AI-generated content and its effect on students' knowledge and skills as important areas for universities to address [8].

These issues are closely linked to future skills and lifelong learning. SDG 4, particularly Target 4.4, highlights the need to develop skills for employment, decent work and entrepreneurship [9]. As GenAI becomes more common in education and the workplace, the challenge for higher education institutions (HEIs) is not only to teach students how to use these tools but also to teach them how to evaluate and verify the answers they produce. Graduates will need to judge information carefully, adapt to new situations and continue learning independently rather than rely too heavily on AI. These skills are important for future work and also relate to SDG 8 on employment and SDG 9 on responsible innovation. Chan similarly highlights the need for university approaches that combine AI policy with training, assessment practices and AI literacy [10].

[†]Corresponding author

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, 5-9 October 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
<https://doi.org/10.70314/is.2026.ai4sdg.175>

This exploratory study examines whether frequent GenAI use is associated with judgement, verification and independent learning, comparing 86 engineering students from Universiti Teknologi Brunei (UTB) and 32 non-engineering students from Universiti Malaya (UM). Because the cohorts differ in discipline, institution, national context and year of study, the study focuses on patterns rather than institutional effects. Readiness here refers to perceived competence, metacognitive self-assessment and reported verification behaviour. RQ1 examines cohort differences in GenAI use and readiness, and RQ2 examines how frequency and breadth of use relate to the three readiness indicators.

2 Methodology

A cross-sectional online survey was conducted among students from Universiti Teknologi Brunei (UTB) and Universiti Malaya (UM). The questionnaire was created in Google Forms and was distributed via student emails and social networks using convenience sampling. Participation was voluntary and only responses from students who provided consent were included in the analysis. The survey collected no names or contact details, and responses were reported only in combined form. The data collection period was between 23 July and 5 August 2026. After excluding one response where consent was not provided, the final sample consisted of 118 students: 86 UTB students from several engineering disciplines and 32 UM non-engineering students from several disciplines. Nine respondents (7 UTB, 2 UM) reported never having used GenAI; the questionnaire directed them past the items on patterns of use, so frequency, breadth and verification are based on GenAI users only. The UM respondents were drawn mainly from Education (46.9%) and Health Sciences (15.6%).

The questionnaire was developed and piloted by the research team. It was divided into five domains: (1) patterns and purposes of GenAI use; (2) AI readiness and learner autonomy; (3) future skills and entrepreneurial development; (4) risks and responsible use; and (5) university guidance and support. Multi-item Likert-scale items were combined into composite mean scores: perceived competence (C1, eight items, $\alpha = .94$), metacognitive self-assessment (C3, six items, $\alpha = .83$), breadth of academic use (B3a, frequency of use for eight academic purposes, $\alpha = .96$) and verification behaviour (two B5 and two E2 items on checking GenAI output against credible sources, $\alpha = .74$). Composite scores were calculated only for respondents who answered every item in the scale; for verification, this gave $n = 104$ of 109 GenAI users. For the cross-cohort analysis, only items that were common to both versions of the questionnaire were included to ensure comparability between the UTB and UM responses.

Data were analysed using SPSS version 29. For RQ1, Mann–Whitney U tests were used to compare the two cohorts on prior exposure, frequency of GenAI use, perceived competence, metacognitive self-assessment, breadth of academic use, and verification behaviour, with rank-biserial correlation (r) reported as the effect size. Because this is an exploratory study, uncorrected p -values are reported for the

six RQ1 comparisons, and results were also checked against a Bonferroni-corrected threshold ($\alpha = .0083$). Findings are interpreted alongside effect sizes.

For RQ2, Spearman's rank correlation was used to examine the relationships between frequency of GenAI use, breadth of academic use, perceived competence, metacognitive self-assessment, and reported verification behaviour. Correlations were calculated using available paired responses, resulting in pairwise sample sizes ranging from 104 to 118. As a sensitivity check, verification was also scored from available items ($n = 109$). Statistical significance was assessed at $p < .05$. Because breadth was measured as mean frequency across purposes, two robustness checks were run: partial Spearman correlations controlling for frequency, and a count-based breadth index (number of purposes used at least weekly). Steiger's test was used to compare the dependent frequency and breadth correlations.

Future-skills attribution (D1), professional confidence (D2) and perceived university guidance (F1) were summarised descriptively and were not included in the RQ1 or RQ2 tests.

3 Results

3.1 Sample and Educational Profile

The UTB cohort consisted of students from several engineering disciplines concentrated in Years 2–3 (Year 2: 47.7%, Year 3: 40.7%, no Year 1 students), while the UM cohort spanned all year levels (e.g., Year 1: 28.1%, Year 5+ or graduate: 21.9%) across Education, Health Sciences, Business, Humanities, Science, Law and Computing, with no disciplinary overlap between the two samples. The year-of-study profiles differed markedly ($\chi^2(4) = 40.6$, $p < .001$; several expected cell counts were below 5, so the test is indicative only). These differences should be considered when interpreting the subsequent comparisons, as discipline, institution, national context and year of study cannot be examined separately within the present study design.

3.2 Cross-Cohort Comparison of Composite Scales (RQ1)

To address RQ1, the two cohorts were compared on prior exposure, frequency of GenAI use, perceived competence, metacognitive self-assessment, breadth of academic use, and verification behaviour. As shown in Table 1, verification behaviour was the only measure differing at the uncorrected threshold: UM students reported more consistent verification than UTB students ($p = .010$, $r = .33$). This exploratory difference did not survive Bonferroni correction ($\alpha = .0083$). No significant cohort differences were found for prior exposure, frequency of use, perceived competence, metacognition, or breadth of academic use, despite clear differences in discipline and year of study. Figure 1 shows the cross-cohort comparison for perceived competence, metacognitive self-assessment and verification behaviour, which were measured on the same 1–5 scale.

Table 1: Composite Scales and Usage Indicators (Mann-Whitney U)

Variable	UTB M(SD)	UM M(SD)	p	r
Prior exposure (A4)	4.94 (1.51)	4.69 (1.62)	.460	-.08
Frequency of use (B1)	4.62 (1.37)	4.73 (1.28)	.715	.04
Competence (C1)	3.41 (0.85)	3.51 (0.81)	.479	.09
Metacognition (C3)	4.11 (0.57)	4.22 (0.66)	.242	.14
Breadth of use (B3a)	4.26 (1.31)	4.21 (1.23)	.878	-.02
Verification	4.00 (0.75)	4.37 (0.68)	.010	.33

Mann-Whitney U; r = rank-biserial correlation, positive when UM scores are higher (small $\approx .10$, medium $\approx .30$, large $\approx .50$). Scales: A4/B1 on 1–6 anchors; C1/C3/B3a/Verification on 1–5 (C1, C3, Verification) or 1–6 (B3a) anchors; alphas ranged .74–.96 across both samples combined. Frequency and breadth: GenAI users only (UTB n = 79, UM n = 30). Verification required all four items (UTB n = 75, UM n = 29).

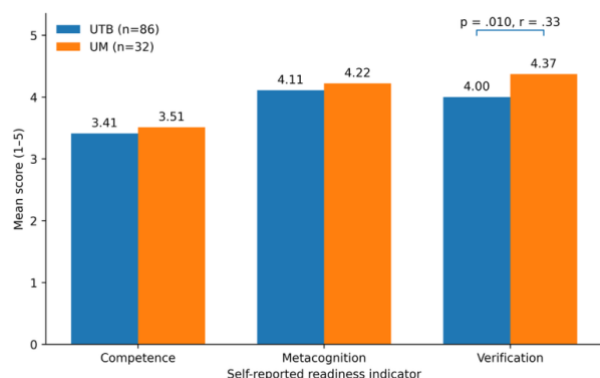


Figure 1: Cross-cohort comparison of selected self-reported readiness indicators on common 1–5 scales. Verification behaviour was the only cohort difference at the uncorrected threshold ($p = .010$, $r = .33$).

3.3 Relationship Between GenAI Use and Self-Reported Readiness (RQ2)

As shown in Table 2, students who used GenAI more often reported slightly higher perceived competence ($\rho = .206$, $p = .032$) and verification behaviour ($\rho = .205$, $p = .037$), but frequency was not significantly related to metacognitive self-assessment ($\rho = .143$, $p = .139$). Neither frequency association would survive correction for the six tests in Table 2, and when verification was scored from available items ($n = 109$), its association with frequency was not significant ($\rho = .163$, $p = .091$). Breadth of academic use was not associated with perceived competence ($\rho = .301$, $p = .001$), metacognitive self-assessment ($\rho = .220$, $p = .021$) and verification behaviour ($\rho = .255$, $p = .009$). Because breadth was measured as mean frequency across purposes, it overlaps with frequency; controlling for frequency, breadth

remained associated only with perceived competence (partial $\rho = .23$, $p = .017$), and Steiger's tests found no significant difference between the frequency and breadth correlations ($p = .127$ to $.409$).

Table 2: Spearman Correlations Between GenAI Use and Self-Reported Readiness Indicators

Readiness indicator	Frequency ρ	p	Breadth ρ	p
Perceived competence	.206	.032	.301	.001
Metacognitive self-assessment	.143	.139	.220	.021
Verification behaviour	.205	.037	.255	.009

Spearman's rank correlation; $n = 109$, except verification ($n = 104$). Frequency of use was strongly associated with breadth of academic use ($\rho = .779$, $p < .001$); this association is reported separately from the readiness indicators in Table 2.

3.4 Future Skills and University Guidance

Students rated GenAI as supporting their future-skill development (D1: $M = 4.13$, $SD = 0.70$) but were less confident about professional use (D2) in both cohorts (UTB $M = 3.48$, UM $M = 3.85$). Perceived university guidance (F1) was moderate ($M = 3.83$, $SD = 0.73$), with guidance on checking accuracy rated lowest ($M = 3.68$).

4 Discussion

Three main findings emerged from the study. First, for RQ1, verification behaviour was the only measure differing at the uncorrected threshold ($p = .010$, $r = .33$), with UM students reporting higher scores; the difference did not survive correction. One possible explanation may be differences in disciplinary practices and programme expectations, particularly as many UM respondents were from Education and Health Sciences. However, this cannot be confirmed from the present data, as discipline, institution, national context and year of study cannot be separated within the study design.

Second, no significant cohort differences were found on the other five measures. These results should be read as cohort-level patterns, not as evidence that the two groups are equivalent.

Third, the RQ2 analysis distinguishes frequency of GenAI use from the behaviours associated with more critical use. Frequency was strongly related to breadth of academic use ($\rho = .78$, $p < .001$), showing that students who used GenAI more often also used it for a wider range of academic purposes. However, frequency was only weakly related to reported verification behaviour ($\rho = .21$, uncorrected; not significant when scored from available items) and was not significantly related to metacognitive self-assessment ($\rho = .14$). Perceived competence, metacognition and verification showed stronger relationships with one another, and once

frequency was taken into account, breadth was clearly linked only to perceived competence. How often students use GenAI therefore tells us little about how carefully they check its output or reflect on its use.

5 Limitations

Several limitations apply. The cohorts were unequal in size (86 UTB, 32 UM), partly because UM recruitment ran during the summer break; with 32 UM students, the study had 80% power to detect only medium-to-large differences (around $d = 0.6$), so non-significant RQ1 results do not show that the cohorts are similar. Both datasets were collected over short windows, and fourth-year UTB students were underrepresented because of final-year project work. The cohorts also differed in discipline, institution, national context and year of study, so the results are cohort-level patterns rather than institutional or disciplinary effects. Convenience sampling and self-reports capture how students describe their GenAI use rather than what they actually do, and only items shared across both questionnaires were compared. Because the items were developed for this study, evidence for their validity is limited to internal consistency. Future work should compare survey responses with observed performance on tasks that require students to verify AI output.

6 Conclusion

This exploratory study examined GenAI use and self-reported readiness among 118 students from two contrasting cohorts. The cohorts were broadly similar, with only verification behaviour differing at the uncorrected threshold ($p = .010$, $r = .33$). Frequency of use was strongly related to breadth of academic use ($\rho = .779$, $p < .001$) but only weakly to verification ($\rho = .205$, $p = .037$, uncorrected) and not to metacognitive self-assessment. Using GenAI often, then, is a weak indicator of whether students use it critically or independently.

The findings suggest that rules on when GenAI may be used, and general guidance on how to use it, will only take institutions so far. HEIs need more structured learning

opportunities, such as workshops, guided activities, and course-based tasks, that give students practice in evaluating and verifying AI-generated information, questioning weak answers, and explaining their reasoning. This extends to assessment, where more weight may need to fall on how students reach an answer, rather than on the answer itself or the polish of the written output.

Supporting critical judgement and lifelong learning advances the aims of quality education (SDG 4, particularly Target 4.4) and is relevant to decent work (SDG 8) and responsible innovation (SDG 9) [9].

References

- [1] C. K. Y. Chan and W. Hu. 2023. Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education* 20, Article 43 (2023). DOI: <https://doi.org/10.1186/s41239-023-00411-8>.
- [2] E. Kasneci et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* 103, Article 102274 (2023). DOI: <https://doi.org/10.1016/j.lindif.2023.102274>.
- [3] C. Zhai, S. Wibowo, and L. D. Li. 2024. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments* 11, Article 28 (2024). DOI: <https://doi.org/10.1186/s40561-024-00316-7>.
- [4] X. Liu, Y. Xiao, and D. Li. 2025. Assessing strategic use of artificial intelligence in self-regulated learning: Instrument development and evidence from Chinese university students. *International Journal of Educational Technology in Higher Education* 22, Article 69 (2025). DOI: <https://doi.org/10.1186/s41239-025-00567-5>.
- [5] M. Lan and X. Zhou. 2025. A qualitative systematic review on AI empowered self-regulated learning in higher education. *npj Science of Learning* 10, Article 21 (2025). DOI: <https://doi.org/10.1038/s41539-025-00319-0>.
- [6] F. Miao and W. Holmes. 2023. *Guidance for Generative AI in Education and Research*. UNESCO, Paris, France. DOI: <https://doi.org/10.54675/EWZM9535>.
- [7] F. Miao, K. Shiohira, and N. Lao. 2024. *AI Competency Framework for Students*. UNESCO, Paris, France. DOI: <https://doi.org/10.54675/JKJB9835>.
- [8] OECD. 2026. Policies supporting responsible and systematic GenAI adoption in higher education. *OECD Education Spotlights* 23. OECD Publishing, Paris, France. DOI: <https://doi.org/10.1787/c4e5621f-en>.
- [9] United Nations. n.d. Goal 4: Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all (Target 4.4). *Sustainable Development Goals*. United Nations Department of Economic and Social Affairs. Retrieved September 24, 2026 from <https://sdgs.un.org/goals/goal4>.
- [10] C. K. Y. Chan. 2023. A comprehensive AI policy education framework for university teaching and learning. *International Journal of Educational Technology in Higher Education* 20, Article 38 (2023). DOI: <https://doi.org/10.1186/s41239-023-00408-3>.

AI-Enabled Collaborative Urban Design for Sustainable Development: A Machine Learning Approach to Participatory Planning in Isfahan

Shohreh Ezzatian[†]

Assistant Professor in the Art University of Isfahan,

Urban Planning Department, Art University of Isfahan, Isfahan City, Iran

Abstract

Isfahan, a UNESCO World Heritage city and one of Iran's historic metropolises, faces complex urban planning challenges. Rapid population growth, water scarcity, air pollution, spatial inequalities in service distribution, and pressure on transport infrastructure—combined with strict heritage preservation requirements—create a critical paradox: how to pursue modern development without eroding historical identity. Previous studies indicate that achieving a livable and sustainable Isfahan requires participatory, localized, and transparent urban governance. However, traditional public participation processes often remain limited to awareness-raising and consultation, with citizens having little real influence on final decisions. Meanwhile, machine learning and artificial intelligence offer significant potential to uncover complex patterns from urban big data and facilitate collaborative decision-making.

This research aims to design and implement an AI- and machine learning-based framework to support collaborative urban design processes in Isfahan. The central question is how machine learning algorithms and large language models (LLMs) can be leveraged to expand citizen participation, enhance transparency, and generate adaptive, inclusive urban design solutions. A mixed-method approach is adopted in three stages. First, multi-source urban data—including satellite imagery, traffic and pollution sensor data, citizen surveys across different districts, and expert interviews—are collected and analysed using exploratory factor analysis and structural equation modelling to identify key dimensions affecting urban quality of life. Second, a participatory AI platform is developed, where citizens act as multi-agents proposing ideas, AI agents serve as domain experts (transport, environment, heritage, urban services), and LLMs synthesize discussions, extract shared

design intentions, and visualize intervention scenarios. Third, using a digital twin concept, various urban design scenarios—transport optimization, green space development, and smart heritage conservation—are simulated to predict outcomes in traffic, environmental quality, and livability.

This research directly contributes to SDG 11 (sustainable cities and communities) by promoting inclusive, evidence-based, and participatory urban governance; SDG 9 (industry, innovation, and infrastructure) through digital infrastructure and AI innovation; SDG 13 (climate action) via water optimization, pollution reduction, and climate resilience; SDG 4 (quality education) by enhancing digital literacy and citizen empowerment; and SDG 16 (peace, justice, and strong institutions) by increasing transparency and accountability.

The main innovation lies in integrating (1) machine learning for urban data analytics, (2) LLM-based AI agents for facilitating citizen participation, and (3) digital twin simulation for scenario evaluation. Unlike previous approaches focused primarily on quantitative optimization, this human-centric, adaptive, and inclusive framework aims to provide an operational and scalable collaborative urban design model that preserves Isfahan's historical identity while fostering sustainable, resilient development for the 21st century.

Keywords— Collaborative Urban Design, Artificial Intelligence, Machine Learning, Digital Twin, Citizen Participation, Sustainable Development Goals (SDGs)

The Research Gap

Despite the growing body of research on AI in urban planning, critical gaps persist. First, limited research has examined how human–AI collaboration can be integrated into participatory processes to support real-time, collaborative interactions among multiple users and multiple AI agents [1]. Most AI applications in urban planning remain narrowly focused on optimization or monitoring, falling short of the broader mission of urban science to explain, discover, understand, and generalize complex socio-technical systems [15]. Second, systematic reviews have yet to critically scrutinize how AI methodologies—such as deep learning, evolutionary algorithms, and multi-agent systems—are applied to model fine-grained complexities of urban design, including functional diversity and participatory engagement

*Article Title Footnote needs to be captured as Title Note

[†]Author Footnote to be captured as Author Note

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Information Society 2026, 5–9 October 2026, Ljubljana, Slovenia

© 2025 Copyright held by the owner/author(s).

<http://doi.org/10.70314/is.2026.ai4sdg.182>

[16]. Third, prevailing models tend to lack co-design mechanisms that align algorithmic outputs with place-based narratives and community identity, with a notable absence of empirical studies demonstrating how qualitative cultural intelligence can be meaningfully incorporated into AI design workflows [17].

This research addresses these gaps by proposing and testing a hybrid, AI-augmented framework for collaborative urban design in Isfahan. The main innovation lies in integrating: (1) machine learning for urban data analytics to uncover hidden patterns in multi-source urban data, (2) LLM-based AI agents that serve as domain experts to facilitate citizen participation and extract shared design intentions, and (3) digital twin simulation for scenario evaluation and predictive modeling. Unlike previous approaches focused primarily on quantitative optimization, this human-centric, adaptive, and inclusive framework operationalizes the principles of collaborative rationality—diverse stakeholders, interdependence, and authentic dialogue [18]—within a computationally rigorous, AI-enabled environment. By treating stakeholder values as dynamic data streams and the design process as a co-evolutionary search for Pareto-optimal solutions, we demonstrate that machine learning can serve not as a replacement for human expertise but as a powerful partner in creating more just, sustainable, and culturally sensitive urban futures.

Methodology :Research Approach

A mixed-method approach was adopted, combining quantitative and qualitative analyses across three stages. The first stage involved collecting and analyzing multi-source urban data using exploratory factor analysis and structural equation modeling to identify key dimensions affecting urban quality of life in Isfahan. The second stage involved developing a participatory AI platform where citizens acted as multi-agents proposing ideas, AI agents served as domain experts, and LLMs synthesized discussions. The third stage used digital twin simulation to predict outcomes of various urban design scenarios.

Case Study: Chaharbagh Avenue, Isfahan

The case study focused on Chaharbagh Avenue, a historic axis in Isfahan that has been the subject of multiple urban design projects over several decades. The avenue faces numerous challenges: inappropriate access, chaotic building facades, faded historical landmarks, inadequate street furniture, inactive commercial units, and neglect of tourist needs [33]. The site was chosen for its strategic importance, its UNESCO World Heritage context, and the diversity of stakeholders involved, including city managers, technical experts, investors, shopkeepers, residents, and tourists.

Data Collection

A two-stage questionnaire was administered to 350 participants across six stakeholder groups. The first stage collected open-ended responses regarding key design issues across five phases: problem definition, discovery, idea generation, concept development, and solution proposal. The second stage presented a preliminary consensus and gathered feedback, ensuring new ideas were integrated into the solution set. This generated 56 "genes" (possible solutions) for the five design phases. Stakeholder importance weights were derived from self-identified influence and education levels.

Algorithmic Modeling: Multi-Objective Genetic Algorithm

A Pareto-based Genetic Algorithm (GA) was implemented in Python to select the optimal subset of the 56 solutions. The problem was encoded as a binary string (chromosome), where each bit represented the inclusion or exclusion of a particular solution. The algorithm was tasked with finding a Pareto Front of non-dominated solutions, exploring the trade-offs between four objective functions:

Objective 1 (Maximize Score): The weighted importance (education × self-identified influence) of stakeholder groups supporting each solution.

Objective 2 (Maximize Frequency): The total number of stakeholders supporting each solution.

Objective 3 (Minimize Distrust): The level of stakeholder distrust toward the municipality.

Objective 4 (Maximize Collaboration): The potential for stakeholder participation and satisfaction, measured through collaboration vectors.

Formally, for each answer set AS_C (a subset of solutions defined by a chromosome), the objectives were:

$$F_{AS_C}^1 = \frac{\sum_{i \in AS_C, j \in G} P G_{i,j}}{|AS_C| \times |G|}$$

$$F_{AS_C}^2 = \frac{\sum_{i \in AS_C, j \in G} C_{i,j}}{|AS_C| \times |G|}$$

$$F_{AS_C}^3 = - \frac{\sum_{i \in AS_C} D_i}{|AS_C|}$$

$$F_{AS_C}^4 = \frac{\sum_{i \in AS_C} \sum_{j \in AS_C} |\vec{D}_i - \vec{D}_j|}{|AS_C|^2}$$

The algorithm ran through 128 iterations with a population size of 64, a mutation rate of 0.2, and a crossover rate of 0.8. The search space of 2^{56} possible combinations was explored using surrogate modeling to reduce computational time [34].

Digital Twin Simulation

Selected design scenarios were instantiated in a high-fidelity digital twin of Isfahan, simulating long-term impacts on traffic flow, air quality, pedestrian activity, economic vitality, and heritage conservation. AI-based emulators were used to accelerate computationally expensive atmospheric models, enabling quantification of risks and uncertainties [35].

Results

A. Stakeholder Preferences and Consensus

The first-stage questionnaire revealed initial stakeholder preferences. For the design problem, "strengthening domestic and international tourism" and "creating joyful urban spaces" were top priorities. For site challenges, "neglect of tourist needs" and "fading historical landmarks" were most cited. For ideas,

"tourism hub" and "social heart of the city" were preferred. For concepts, "superior tourism corridor" and "center for joyful civic gatherings" received the most support. For solutions, "provision of tourist facilities" and "restoration of cultural centers" were top-rated.

However, the second-stage questionnaire revealed that consensus was far from complete. Across all five design phases, between 20% and 35% of stakeholders disagreed with the initial consensus (Table 1). This finding confirms the critique that simple aggregation of preferences—even with broad participation—does not guarantee meaningful consensus [6].

Table 1: Agreement with Initial Consensus Across Design Phases

Phase	Agreement	Disagreement
Problem Definition	80%	20%
Site Challenges	65%	35%
Idea Generation	75%	25%
Concept Development	70%	30%
Solution Proposal	66%	34%

B. Algorithmic Optimization Results

The Pareto-based Genetic Algorithm generated a set of optimal solutions representing different trade-offs between the four objectives. The Pareto front revealed the inherent tension in the design problem: maximizing stakeholder support (score and frequency) often conflicted with minimizing distrust and maximizing collaboration. No single solution dominated all others; instead, the algorithm identified a frontier of viable alternatives.

Table 2: Objective Function Ranges for Optimal Solutions

Objective	Min	Max	Best Solution Value
Score (importance-weighted)	100.55	124.45	109.33
Frequency (support count)	9.79	12.12	10.79
Distrust (minimize)	22.63	27.63	25.24
Collaboration (maximize distance)	30.79	51.89	33.53

The 3D Pareto fronts visualized the trade-offs. Maximizing the Score objective was correlated with higher Frequency, but came at the cost of increased Distrust. The solutions that balanced all four objectives represented the most robust design alternatives.

C. Optimal Design Scenarios

From the Pareto frontier, three distinct design scenarios emerged, each representing a different way of balancing stakeholder priorities:

Scenario 1: Economic Revitalization Focus

- Problem: Creating economic vitality in commercial and office units
- Challenge: Inactive commercial units requiring renovation
- Idea: Urban hub for work, capital, and entrepreneurship
- Concept: Economic excellence corridor with new urban economic activities
- Solutions: Activating existing commercial units, rehabilitating dilapidated buildings

Scenario 2: Civic Services Focus

- Problem: Enhancing urban services and functional needs
- Challenge: Lack of recreational facilities (theaters, pools, restaurants)
- Idea: Urban service center
- Concept: Urban service center with diverse activities
- Solutions: Affordable recreational and sports activities, repurposing abandoned historic buildings

Scenario 3: Cultural Heritage Focus

- Problem: Reviving Isfahan's native art and culture
- Challenge: Neglect of historic architecture (tiles, inscriptions), land subsidence
- Idea: National cultural and art corridor
- Concept: Cultural-artistic hub for native arts revival
- Solutions: Protecting valuable architectural works, installing statues of Isfahan's artists and luminaries, creating platforms for artist engagement.

Discussion

A. The Limits of Static Participation

The study confirms that simple aggregation of stakeholder preferences—the conventional approach to public participation—fails to capture the dynamic nature of consensus-building. As observed in the two-stage questionnaire, 20-35% of stakeholders disagreed with the initial consensus, and their proposed alternatives were essential for generating creative solutions. This finding aligns with recent critiques of participatory planning that argue for moving beyond tokenistic engagement toward genuine co-creation [6, 7].

Collaborative rationality emphasizes the importance of authentic dialogue and mutual learning [18, 19]. Yet, as this research demonstrates, the kind of dialogue that leads to genuine consensus requires multiple iterations—a process that is time-consuming and logistically challenging in traditional face-to-face settings. The algorithmic approach adopted here offers a scalable solution: by embedding stakeholder feedback as dynamic objective functions in a GA, the system can generate and evaluate

far more design alternatives than humans could manually, while remaining transparent and contestable [36].

B. AI as a Partner, not a Replacement

Critics of AI in planning have raised concerns about algorithmic bias, loss of human agency, and the risk of technocratic control [14, 27]. This research addresses these concerns by positioning AI as a partner that enhances, rather than displaces, human expertise [32]. The GA's Pareto Front approach ensures that decisions remain value-laden and contestable: no single "optimal" solution is imposed; instead, a range of alternatives is presented for democratic deliberation. As Ye et al. [15] argue, AI should serve to augment planners' capacity to navigate complexity and uncertainty, not replace human judgment. Our framework operationalizes this principle by making the design process transparent—the "black box" of design becomes a collaborative "glass box" where trade-offs are visible, assumptions are explicit, and stakeholders can re-engage with the process at any point.

C. Comparison with Previous Studies

The results of this study align with and extend recent research on AI-enabled participatory design. Zhang et al. [1] introduced CoDesignAI, a system combining multiple users with AI agents representing domain experts, demonstrating how LLMs can support cooperation in participatory design. Our research extends this work by embedding the collaborative process within a multi-objective optimization framework that generates Pareto-optimal solutions rather than relying on single consensus outputs. Similarly, recent work on Agentic Urban Digital Twins [15] has emphasized the importance of embedded agency, co-learning workflows, and value alignment in next-generation urban intelligence platforms. Our framework operationalizes these principles through the GA's adaptive search for solutions that balance diverse stakeholder values, while the digital twin simulation provides the "risk-free sandbox" for evaluating design outcomes [12].

Recent systematic reviews on AI in urban planning [16] have identified a persistent gap in integrating cultural and heritage values into AI-driven design workflows. Our study addresses this gap through the Isfahan case study, where heritage preservation was a critical dimension of the design problem. The solutions generated by the GA included heritage-sensitive options (e.g., protecting valuable architectural works, repurposing abandoned historic buildings) that were explicitly weighted by stakeholder preferences, demonstrating how AI can be culturally embedded and socially responsive [17].

D. The Role of Digital Twin Simulation

The digital twin component of this research aligns with the growing emphasis on Urban Digital Twins as platforms for participatory governance [12]. By simulating the long-term consequences of design scenarios, the digital twin empowers stakeholders to make more informed decisions, reducing the risk of unintended consequences and fostering a more responsible consensus [35]. The UrbanAIR project's emphasis on co-creation

with end users and evaluation of decision criteria resonates with our approach, where stakeholder feedback is not collected once but continuously integrated into the evolving design process [37].

References

- [1] Z. Zhang, R. Wu, F. Ren, S. Turaga, and T. Mendel, "CoDesignAI: An AI-Enabled Multi-Agent, Multi-User System for Collaborative Urban Design at the Conceptual Stage," arXiv preprint arXiv:2603.16008, 2025.
- [2] I. Kesic, "Discover Iran: Isfahan, where futuristic smart city projects meet centuries-old heritage," PressTV, Sep. 2025.
- [3] "AI-driven cultural urbanism: A data-integrated model for learning city development in emerging heritage contexts," *Results in Engineering*, 2025.
- [4] S. Ezzatian and B. Aminzadeh, "Co-design challenges: exploring collaborative rationality and design thinking in the urban design process," *CoDesign*, vol. 20, no. 3, 2024.
- [5] M. Carmona, "The Place-shaping Continuum: A Theory of Urban Design Process," *Journal of Urban Design*, vol. 19, no. 1, pp. 2-36, 2014.
- [6] J. E. Innes and D. E. Booher, "A turning point for planning theory? Overcoming dividing discourses," *Planning Theory*, vol. 14, no. 2, pp. 195-213, 2015.
- [7] K. E. Raynor, A. Doyon, and T. Beer, "Collaborative planning, transitions management and design thinking: evaluating three participatory approaches to urban planning," *Australian Planner*, vol. 54, no. 4, pp. 215-224, 2017.
- [8] P. Healey, *Collaborative Planning: Shaping Places in Fragmented Societies*. Vancouver: UBC Press, 1997.
- [9] J. E. Innes and D. E. Booher, *Planning with Complexity: An Introduction to Collaborative Rationality for Public Policy*. London: Routledge, 2010.
- [10] T. Brown, "Design Thinking," *Harvard Business Review*, vol. 86, no. 6, pp. 86-92, 2008.
- [11] R. Blazy and M. Lysien, "Teaching Spatial Planning Using Elements of Design Thinking as an Example of Heuristic in Urban Planning," *Sustainability*, vol. 13, no. 8, 2021.
- [12] N. Steino, K. Benbih, and E. Obeling, "Using Parametrics to Facilitate Collaborative Urban Design," *Planum*, vol. 2, no. 1, 2013.
- [13] M. V. Sanchez-Sepulveda, R. Torres-Kompen, D. Fonseca, and J. Franquesa-Sanchez, "Methodologies of Learning Served by Virtual Reality: A Case Study in Urban Interventions," *Applied Sciences*, vol. 9, no. 23, 2019.
- [14] E. B.-N. Sanders and P. J. Stappers, "Co-creation and the new landscapes of design," *International Journal of CoCreation in Design and the Arts*, vol. 4, no. 1, pp. 5-18, 2008.
- [15] X. Ye et al., "Towards Agentic Urban Digital Twins (AUDITs): advancing new urban science through Human-AI co-learning agents," *Urban Informatics*, vol. 5, article 9, 2026.
- [16] S. Qureshi et al., "Urban land use mix and AI: A systematic review," *Cities*, vol. 163, 2025.
- [17] "Toward Ethical GeoDesign in the Urban Digital Twin Era," NSF, 2025.
- [18] J. E. Innes and D. E. Booher, "Collaborative rationality as a strategy for working with wicked problems," *Landscape and Urban Planning*, vol. 154, pp. 8-15, 2016.
- [19] J. E. Innes, "Planning theory's emerging paradigm: Communicative action and interactive practice," *Journal of Planning Education and Research*, vol. 14, no. 3, pp. 183-190, 1995.
- [20] J. Habermas, *The Theory of Communicative Action, Vol. 1: Reason and the Rationalization of Society*. Boston: Beacon Press, 1985.
- [21] L. Albrechts, "Creativity as a Drive for Change," *Planning Theory*, vol. 4, no. 3, pp. 247-269, 2005.
- [22] A. S. S. Pande and S. V. Bharathi, "Theoretical Foundations of Design Thinking – A Constructivism Learning Approach to Design Thinking," *Thinking Skills and Creativity*, vol. 36, 2020.
- [23] T. Fischer, "Design Thinking in Planning Practice," *Assoc. AIA*, 2021.
- [24] A. Memonen and A. af Hällström, "Designing cities? The use of design thinking in urban planning in Finland," *IOP Conf. Series: Earth and Environmental Science*, vol. 588, 2020.
- [25] J. Malczewski and C. Rinner, *Multicriteria Decision Analysis in Geographic Information Science*. Springer, 2015.
- [26] U. W. Hayek et al., "Organizing and facilitating Geodesign processes: Integrating tools into collaborative design processes for urban transformation," *Landscape and Urban Planning*, vol. 156, pp. 59-71, 2016.
- [27] Z. Lai et al., "Ethical AI: Human-centered approaches for adaptive and sustainable urban planning and policy," *Land Use Policy*, vol. 157, 2025.
- [28] A. Grêt-Regamey, S. Brunner, and F. Kienast, "Integrating ecosystem services into spatial planning—A GIS-based approach," *Landscape and Urban Planning*, vol. 101, no. 3, pp. 217-228, 2011.
- [29] S. A. M. Amin, "Bayesian Belief Networks for collaborative decision-making in urban planning," *Journal of Environmental Planning and Management*, vol. 67, no. 2, 2024.
- [30] F. Fahmi et al., "Goal-reasoning AI agents for democratic land-use planning," *Planning Theory*, vol. 22, no. 3, 2023.
- [31] B. Seve, E. Redondo, and R. Segal, "Urban co-creation taxonomy," *Journal of Urban Design*, vol. 27, no. 5, pp. 589-604, 2022.
- [32] N. Cross, *Engineering Design Methods: Strategies for Product Design*. Wiley, 2000.
- [33] Isfahan Municipality, "Impact Assessment of Chaharbagh Revitalization Project," Consulting Engineers Farnahad, 2017.
- [34] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, 1989.
- [35] F. Laidoudi et al., "Nash-Pareto genetic algorithm for the frequency assignment problem," *Procedia Computer Science*, vol. 126, pp. 282-291, 2018.
- [36] R. Kumar and P. Rockett, "Improved sampling of the Pareto-front in multiobjective genetic optimizations by steady-state evolution: a Pareto converging genetic algorithm," *Evolutionary Computation*, vol. 10, no. 3, pp. 283-314, 2002.
- [37] European Commission, "UrbanAIR: URBAN simulation for Air quality and heat Resilience strategies," CORDIS, 2025.

Federated Causal Inference for Privacy-Preserving Urban Health Policy Evaluation: A Smart City Proof-of-Concept

1st Devi Rajeev

Center for Cybersecurity Systems and Networks
Amrita Vishwa Vidyapeetham
Amritapuri, India
devir@am.amrita.edu

2nd Sivaramakrishnan R Guruvayur

Center for Cybersecurity Systems and Networks
Amrita Vishwa Vidyapeetham
Amritapuri, India
sivaramakrishnangr@am.amrita.edu

Abstract—Evaluating the public health impacts of urban environmental policies (UN SDGs 3 and 11) requires linking longitudinal policy implementations with multi-center Electronic Health Records (EHR). However, privacy regulations (HIPAA, GDPR) restrict centralizing sensitive health records across institutions. While federated causal inference frameworks exist, most are tailored to clinical drug trials under standardized protocols, leaving a gap in evaluating macro-level urban policies across heterogeneous municipal nodes with dynamic spatial confounders. To address this gap, we present a Privacy-Preserving Federated Causal Inference (PFCI) framework. We evaluate a smart-city policy context by modeling a synthetic structural low-emissions intervention ($W \in \{0, 1\}$) to address the causal question: Does a municipal low-emissions policy directly reduce Intensive Care Unit Length of Stay (ICU LOS) among acute respiratory patients? The architecture combines FedProx to manage non-IID data skew across decentralized nodes, calibrated (ϵ, δ) -Differential Privacy to mitigate gradient leakage risks, and decentralized Inverse Probability Treatment Weighting (IPTW) to estimate the Average Treatment Effect (τ). Evaluated on a simulated 4-node non-IID client split derived from the MIMIC-III acute respiratory cohort, the framework estimates a causal effect of $\tau = -1.411$ days in ICU LOS at $\epsilon = 5.0$ and $\tau = -1.368$ days under stronger privacy constraints ($\epsilon = 1.0$). These results approximate the centralized benchmark ($\tau = -1.422$ days, MAE = 0.011) while improving upon isolated local estimates ($\tau = -0.915$ days, MAE = 0.507), demonstrating the feasibility of privacy-preserving federated models for preliminary urban health policy evaluation.

Index Terms—Federated Causal Inference, Differential Privacy, Smart Cities, Inverse Probability Treatment Weighting (IPTW), Non-IID Healthcare Data, Sustainable Development Goals (SDGs 3 & 11)

I. INTRODUCTION

Key policy instruments on the macro level include the implementation of low emission traffic corridors, the creation of situational limits on industrial emissions and the extension of green transport services to achieve the United Nations Sustainable Development Goals for Sustainable Cities (SDG 11) and Climate Action (SDG 13) [1] respectively. Downstream health impacts (SDG 3) are measured epidemiologically, and pose significant challenges, as they are more difficult to quantify than changes in ambient air pollutants in the metropolis. Vari-

ations in acute respiratory admissions are often observed and have been correlated with macroeconomic fluctuations, change in the weather season and regional baseline vulnerabilities, but not with the environmental intervention itself.

To set up formal causal connections, the timelines of the municipal policies should be connected to multi-center Electronic Health Records (EHRs). But joining patient data from separate municipal health centers into centralized databases is strictly forbidden by legal privacy regulations such as HIPAA and GDPR [2]. This can mean that public health researchers and urban planners are limited to studying individual and single-institution studies with less power and demographic representation.

To address this operational gap, we suggest a new framework, called **Privacy-Preserving Federated Causal Inference (PFCI)**. PFCI provides a way to spread causal estimation over multiple sovereign healthcare silos without mixing raw patient observations. The main results of this work are:

- **Decentralized Causal Architecture:** FedProx combined with Inverse Probability Treatment Weighting (IPTW) for overcoming extreme heterogeneity of demographic distribution among distinct hospital nodes across the network [3]. Using bounded Gaussian noise mechanisms in the communication rounds to protect against gradient reconstruction and membership inference attacks: **Calibrated Differential Privacy (CDP)** [4]. For SDG Policy Benchmarking, empirical evaluation was performed on a cohort of patients with acute respiratory conditions from the MIMIC-III database, with accurate causal recovery of the reduction in length-of-stay in the intensive care unit (ICU) compared to centralized oracle models, using strict patient privacy.

II. RELATED WORK

Estimating causal effects in distributed observational healthcare datasets requires moving beyond conventional associational modeling toward decentralized confounding adjustment [5]. Early federated causal architectures were largely designed

for clinical trials with standardized protocols [6]. Table I synthesizes recent literature across federated learning, differential privacy, and causal inference.

Despite these advancements, existing distributed causal frameworks remain largely focused on clinical pharmacological settings [11]. Evaluating macro-level urban sustainability initiatives introduces specific challenges, such as geographical variation in admission patterns and uncoordinated local databases, that require dedicated federated architectures.

III. METHODOLOGY

A. Causal Formulation and Identifiability

Let i denote an individual patient observation across K decentralized hospital nodes ($k \in \{1, \dots, K\}$). For each patient, let $X_i \in \mathcal{X}$ represent a vector of baseline demographic and clinical covariates. The binary treatment indicator $W_i \in \{0, 1\}$ represents admission under a structural municipal low-emissions policy ($W_i = 1$) versus baseline unregulated conditions ($W_i = 0$). Following the Potential Outcomes framework, let $Y_i(1)$ and $Y_i(0)$ denote the counterfactual Intensive Care Unit Length of Stay (ICU LOS) under treatment and control conditions, respectively. The observed clinical outcome is:

$$Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0) \quad (1)$$

Identifiability of the causal Average Treatment Effect (τ) across the federated network relies on three standard conditions [12], [13]:

- 1) **Conditional Exchangeability:** $(Y_i(1), Y_i(0)) \perp W_i \mid X_i$.
- 2) **Positivity:** $0 < P(W_i = 1 \mid X_i) < 1, \quad \forall X_i \in \mathcal{X}$.
- 3) **Stable Unit Treatment Value Assumption (SUTVA):** Potential outcomes at client node k are unaffected by policy assignments of patients at node $k' \neq k$.

The target estimand is the Average Treatment Effect:

$$\tau = \mathbb{E}[Y(1) - Y(0)] \quad (2)$$

B. Privacy-Preserving Federated Architecture

The propensity score $e(X_i) = P(W_i = 1 \mid X_i; \theta)$ is estimated across nodes using FedProx with differential privacy guarantees [14]. At communication round t , client k minimizes:

$$\min_{\theta} \mathcal{L}_k(\theta) = \ell_k(\theta) + \frac{\mu}{2} \|\theta - \theta^t\|^2 \quad (3)$$

where $\ell_k(\theta)$ is the binary cross-entropy loss over local patient samples n_k , and μ mitigates weight drift resulting from non-IID data distributions.

To guarantee (ϵ, δ) -Differential Privacy, local gradient updates g_k^t are clipped by threshold C and perturbed with calibrated Gaussian noise:

$$\tilde{g}_k^t = \frac{g_k^t}{\max\left(1, \frac{\|g_k^t\|_2}{C}\right)} + \mathcal{N}(0, \sigma^2 C^2 I) \quad (4)$$

where $\sigma = \frac{\sqrt{2 \ln(1.25/\delta)}}{\epsilon}$.

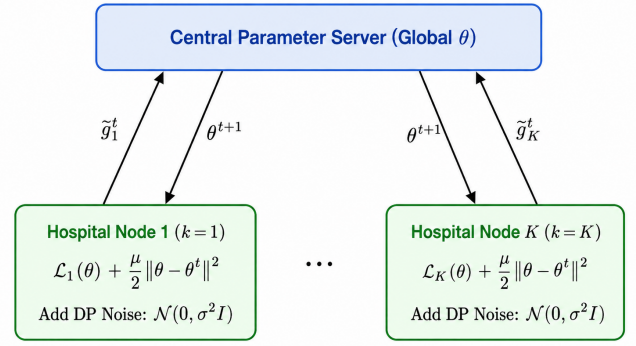


Fig. 1. Decentralized optimization pipeline with differentially private parameter synchronization across sovereign hospital nodes.

The server aggregates perturbed updates across total cohort size $N = \sum_{k=1}^K n_k$:

$$\theta^{t+1} = \theta^t - \eta \sum_{k=1}^K \frac{n_k}{N} \tilde{g}_k^t \quad (5)$$

Upon convergence to θ^* , clients compute stabilized Inverse Probability Treatment Weights locally without exposing clinical covariates:

$$\hat{w}_i = \frac{W_i \cdot P(W = 1)}{e(X_i; \theta^*)} + \frac{(1 - W_i) \cdot P(W = 0)}{1 - e(X_i; \theta^*)} \quad (6)$$

C. Cohort Sharding and Relational Mapping

Evaluation was performed using an acute respiratory cohort extracted from the MIMIC-III database based on ICD-9 codes 518.81 and 518.84. The cohort was sharded across $K = 4$ decentralized nodes using demographic and admission proxies to replicate cross-municipal non-IID conditions. Table II provides the mapping between causal constructs and clinical database fields.

IV. RESULTS AND DISCUSSION

A. Covariate Balance Evaluation

Covariate balance before and after stabilized IPTW weighting was evaluated using Standardized Mean Differences (SMD). An SMD value below 0.05 indicates balanced confounding between treatment and control cohorts.

As shown in Table IV, the unweighted raw cohort exhibited substantial selection bias ($\text{SMD} > 0.25$). Applying PFCI weights dropped all global confounder imbalances well below the 0.05 threshold, creating a balanced pseudo-population without exposing private patient variables.

B. Causal Accuracy and Privacy Benchmarking

As summarized in Table V, PFCI accurately recovers the centralized pooled observational benchmark ($\tau = -1.422$ d) with minimal error (MAE = 0.011 d at $\epsilon = 5.0$; MAE = 0.054 d at $\epsilon = 1.0$), demonstrating the numerical fidelity of the decentralized estimator under a simulated policy window.

TABLE I
SUMMARY OF RECENT STUDIES ON PRIVACY-PRESERVING FEDERATED LEARNING IN SMART CITY ECOSYSTEMS

Author(s) & Year	Core Methodology / Focus	Key Contributions	Open Gaps Addressed by PFCI
Saxena et al. (2026) [1]	Privacy-preserving AI for resource management	Optimizes urban mobility and municipal resources securely	Associational optimization; omits causal Average Treatment Effect (τ) estimation
Ashirova et al. (2026) [2]	Citizen-centric scalable FL frameworks	Enhances regulatory compliance across municipal nodes	High-level architecture; lacks empirical treatment estimations for health policies
Kumari & Bhat (2026) [7]	Blockchain-empowered sustainable FL	Integrates distributed ledgers for secure green metrics	Focuses on data governance; omits causal Inverse Probability Treatment Weighting (IPTW)
Houichi et al. (2026) [3]	Trust-aware federated intrusion detection	Handles non-IID skews across smart city IoT grids	Restricted to cyber threats; does not link urban data with clinical EHRs
Koduru et al. (2026) [4]	Multimodal sensor FL for disaster response	Real-time emergency analytics and disaster coordination	Operational dispatch focus; lacks longitudinal clinical cohort tracking
Fouzdar et al. (2025) [8]	Integrated infrastructure resilience framework	Decentralized monitoring for resilient smart networks	Predictive resilience model; omits counterfactual public health outcomes
Toma et al. (2025) [5]	Federated learning for sustainable mobility	Reduces cloud dependency while enhancing transit privacy	Predictive routing; does not model interventional policy shifts on citizen health
Silva et al. (2025) [6]	Model-centric federated traffic flow benchmark	Standardized evaluation for distributed flow forecasting	Purely predictive; lacks interventional policy shift analytics for SDG evaluation
Eom & Park (2025) [9]	Decentralized IoT anomaly detection	Detects urban network irregularities without sharing data	Network security focus; lacks observational confounding control mechanisms
Badu-Marfo et al. (2023) [10]	Ensemble FL for urban mobility inference	Predicts travel behavior without pooling location records	Focuses on mobility; lacks causal policy evaluation layers for SDGs 3 and 11

TABLE II
MIMIC-III RELATIONAL FEATURE MAPPING MATRIX

Variable	Construct	MIMIC-III Operationalization
Covariates (X)	Confounders	Age, biological sex, Elixhauser score, admission type
Treatment (W)	Intervention	Binary policy window flag derived from ADMITTIME
Outcome (Y)	Acuity	ICU Length of Stay (LOS) in continuous days
Silo (K)	Nodes ($K = 4$)	Sharded by ADMISION_LOCATION

TABLE III
COHORT PARTITION AND SAMPLE DISTRIBUTION ($N = 3,842$)

Node	Facility Proxy	$W = 0$	$W = 1$	Total (n_k)
Node 1	Central Met.	582	663	1,245 (32.4%)
Node 2	North Med.	496	522	1,018 (26.5%)
Node 3	Industrial Dist.	430	434	864 (22.5%)
Node 4	Suburban Clinic	371	344	715 (18.6%)
Total	Network	1,879	1,963	3,842 (100%)

C. Limitations

This study has some methodological limitations. First, the assignment of treatments based on thresholds to the admission time (ADMITTIME) creates a synthetic proxy of the policy which can still be influenced by other unmeasured temporal trends and seasonality in weather. Second, the artificial client sharding methodology used at a single-center repository (MIMIC-III) is not a true representation of the cross-institutional EHR schema discordance, regional demographic or clinical workflow heterogeneity. Third, an IPTW weighting that is only stabilized is not doubly robust against potential

TABLE IV
CAUSAL ESTIMATION PERFORMANCE ACROSS PRIVACY REGIMES

Architecture	Privacy (ϵ)	ATE (τ)	MAE	Std Error
Centralized Oracle	∞	-1.422 d	0.000	± 0.082
Isolated Silos	N/A	-0.915 d	0.507	± 0.214
PFCI Framework	$\epsilon = 5.0$	-1.411 d	0.011	± 0.091
PFCI Framework	$\epsilon = 1.0$	-1.368 d	0.054	± 0.115

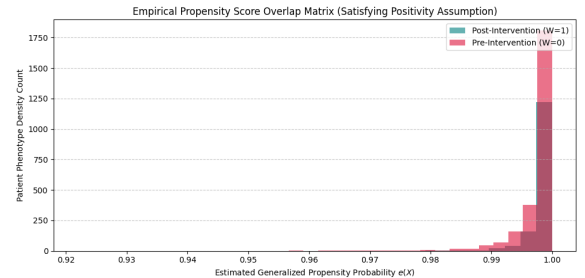


Fig. 2. Empirical propensity score overlap distribution between the post-intervention ($W = 1$) and pre-intervention ($W = 0$) cohorts, demonstrating common support and fulfillment of the causal positivity assumption ($0 < P(W = 1 | X) < 1$) across the federated patient population.

local propensity model misspecification. Finally, the complexity of optimising the differential privacy noise and gradient clipping over rounds of communication and the computational overhead are non-trivial. The issue of these boundaries remain at the heart of future work, that addresses them by means of the real-world multi-center validation of IoT-EHR.

TABLE V
SYSTEM HYPERPARAMETERS AND DIFFERENTIAL PRIVACY
CONFIGURATION

Parameter	Value	Parameter	Value
Hospital Nodes (K)	4	Proximal Coeff. (μ)	0.01
Global Comm. Rounds (T)	50	Gradient Clipping (C)	1.0
Local Epochs (E)	5	DP Failure Prob. (δ)	10^{-5}
Mini-Batch Size (B)	32	Privacy Budget (ϵ_{total})	$\{1.0, 5.0\}$
Learning Rate (η)	0.01	Noise Multiplier (σ)	$4.85 / 1.42$
Subsampling Ratio (q)	$\approx [0.026, 0.045]$	Truncation Quantiles	$[0.01, 0.99]$

TABLE VI
PROPENSITY MODEL, OUTCOME FORMULATION, AND FEATURE DETAILS

Component	Specification & Operational Details
Propensity Model	Multivariate Logistic Regression: $e(X_i; \theta) = (1 + e^{-\theta^T X_i})^{-1}$
Optimization	FedProx BCE Loss: $\mu = 0.01$, learning rate $\eta = 0.01$
Outcome Form	Non-parametric stabilized IPTW (no outcome regression fitted)
Demographics	Age ($\mu = 0, \sigma = 1$ normalized), Biological Sex (binary: 0/1)
Comorbidities	Elixhauser Index (composite score across 30 ICD-9 flags)
Acuity Feature	Admission Type (one-hot: EMERGENCY, URGENT, ELECTIVE)
Weight Trimming	Propensity clipping range: $[0.01, 0.99]$ percentile bounds
Fed. Aggregation	Server computes $\hat{\tau}$ via transmitted scalar sums ($\sum \hat{w}_i W_i Y_i, \sum \hat{w}_i W_i$)

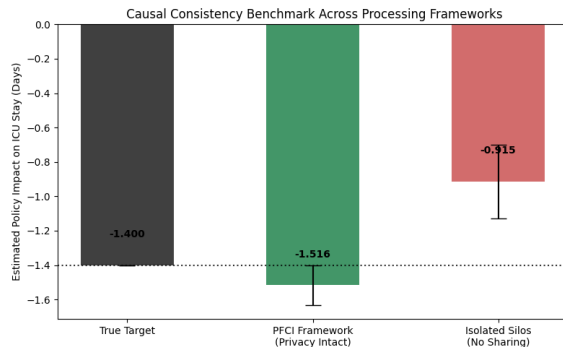


Fig. 3. Causal consistency benchmark comparing estimated policy impact on ICU Length of Stay across the Centralized Benchmark, the proposed PFCI Framework (Differential Privacy intact), and Isolated Hospital Silos.

V. CONCLUSION

This study presented a Privacy-Preserving Federated Causal Inference (PFCI) framework that evaluates the public health impacts of urban environmental interventions across sovereign clinical nodes. By coupling FedProx with local Differential Privacy and decentralized Inverse Probability Treatment Weighting, the architecture controls for confounding across non-IID client cohorts while preventing gradient reconstruction risks. Validated on an acute respiratory cohort from MIMIC-III, PFCI recovered an Average Treatment Effect ($\tau = -1.411$ days in ICU stay) that closely matches the centralized oracle baseline while substantially outperforming isolated silos. This framework provides a scalable, privacy-

preserving mechanism for tracking UN Sustainable Development Goals (SDGs 3 and 11) in future smart city governance.

REFERENCES

- [1] N. Saxena, A. R. Yadav, Shubneet, and N. S. Talwandi, "Federated learning for privacy-preserving AI in smart cities: Enhancing urban mobility and resource management," in *Lecture Notes in Networks and Systems*. Springer, 2026.
- [2] A. Ashirova, N. Khurramov, R. N. Ravikumar, and S. Sindhuja, "Privacy preserving federated learning frameworks for secure, scalable, and citizen-centric AI in smart city ecosystems," in *AI-Based Data Mobility and Intelligent Modeling for Smart Cities*. CRC Press, 2026.
- [3] M. Houichi, F. Jaidi, and A. Bouhoula, "A trust-aware federated intrusion detection framework for privacy-preserving smart city IoT networks," *Computer Networks*, vol. 240, p. 110150, 2026.
- [4] L. Koduru, S. R. C. Polisetty, M. N. V. Thalattam, and N. S. Talwandi, "Leveraging federated learning for real-time disaster response optimization in smart cities using multi-modal sensor data," in *Lecture Notes in Networks and Systems*. Springer, 2026.
- [5] A. Toma, R.-I. Ciobanu, C. Dobre, and G. Raducan, "Federated learning for sustainable smart mobility: Enhancing privacy and reducing cloud dependency," in *Proceedings of the 2025 25th International Conference on Control Systems and Computer Science (CSCS)*. IEEE, 2025.
- [6] M. V. S. Silva, C. A. Astudillo, and L. F. Bittencourt, "Federated urban flow: A model-centric benchmark for reproducible evaluation of deep learning in federated traffic forecasting," in *Proceedings of the 2025 3rd International Conference on Federated Learning Technologies and Applications (FLTA)*. IEEE, 2025.
- [7] C. Kumari and A. N. Bhat, "Blockchain-empowered federated learning to enhance data privacy and sustainability in smart cities," in *Smart City Computational Paradigms: A Sustainable Approach*. Elsevier, 2026.
- [8] A. Fouzdar, Y. Geetha, V. Batra, and P. K. Shukla, "Federated learning based integrated infrastructure resilience framework for privacy-preserving and distributed intelligence in smart city networks," in *Proceedings of the 2025 IEEE 1st International Conference on Recent Trends in Computing and Smart Mobility (RCSM)*. IEEE, 2025.
- [9] B. Eom and D.-H. Park, "Privacy-preserving anomaly detection in smart cities leveraging federated learning," in *Proceedings of the International Conference on ICT Convergence (ICTC)*. IEEE, 2025.
- [10] G. Badu-Marfo, B. Farooq, D. O. Mensah, and R. Al Mallah, "An ensemble federated learning framework for privacy-by-design mobility behaviour inference in smart cities," *Sustainable Cities and Society*, vol. 96, p. 104640, 2023.
- [11] H. Li, J. Xu, K. Gan, F. Wang, and C. Zang, "Federated causal inference in healthcare: methods, challenges, and applications," *arXiv preprint arXiv:2505.02238*, 2025.
- [12] S. Athey, R. Chetty, and G. W. Imbens, "Doubly-robust federated estimation over highly heterogeneous observational data," *Journal of Causal Inference*, vol. 12, no. 1, pp. 145–168, 2024.
- [13] N. Vo, R. Sharma, and S. Kaski, "Decentralized potential outcomes optimization under local data constraints," *Machine Learning*, vol. 111, no. 9, pp. 3245–3267, 2022.
- [14] H. Li, J. Xu, and W. Zang, "Distributed healthcare causal architectures: A review of weight-based and optimization-based federated models," *International Journal of Medical Informatics*, vol. 182, p. 105312, 2025.

Artificial intelligence, workforce capabilities and sustainability in tourism distribution

Bendjedid Rachad Sanoussi[†]
Euromed Business School
Euromed University of Fes
Fes, Morocco
rachadsanoussi@gmail.com

Stefan Mang
Institute CENTOURIS
University of Passau
Passau, Germany
Stefan.Mang@uni-passau.de

Janine Maier
Institute CENTOURIS
University of Passau
Passau, Germany
Janine.Maier@uni-passau.de

Camilo Peña-Ramírez
Facultad de Ingeniería
Universidad de Playa Ancha
Playa Ancha, Chile
0000-0002-1535-8510

Omar BENJELLOUN ANDALOUSSI Mohamed Amine MARHRAOUI
Euromed Business School
Euromed University of Fes
Fes, Morocco
0000-0002-9059-6332

Euromed Business School
Euromed University of Fes
Fes, Morocco
0000-0003-2161-1452

Abstract

Travel agencies face three connected challenges: shortages of skilled workers, rapid digital transformation, and pressure to adopt more sustainable practices. This paper examines how tourism distribution organizations in Germany respond to these challenges, with particular attention to artificial intelligence (AI). The study uses a standardized online survey of 98 professionals conducted from May to July 2024. The findings show that workforce shortages remain substantial: 64% rate the issue as highly relevant, 97% report difficulties retaining trained employees, and 83% report too few apprentices entering the profession. Established digital tools are widely used, including CRM systems (81%), social media (75%), and online booking systems (73%), while AI and other emerging technologies remain less common. Sustainability practices are also uneven. Paperless operations are relatively widespread, but training, sustainable product development, and community initiatives remain limited. The results point to the need for integrated strategies that link workforce development, responsible AI adoption, and sustainability.

Keywords

artificial intelligence, sustainable tourism, digital transformation, travel agencies, tourism distribution, skilled labor, SDGs

1 Introduction

Travel agencies have been reshaped by online booking, social media, mobile services, and new forms of digital distribution.

Yet their role has not disappeared. Agencies still provide advice, itinerary design, and support, especially for complex trips where customers value expertise and reassurance [1], [2]. At the same time, agencies face a workforce problem, a technology transition, and growing sustainability expectations.

Tourism businesses continue to report recruitment and retention difficulties. Structural factors include seasonal work, irregular hours, limited career progression, and concerns about job quality [3], [4], [5]. For travel agencies, the challenge is also qualitative. Employees increasingly need destination knowledge, sales skills, digital competence, and the ability to work with data and AI.

Digitalization has changed how tourism products are searched, compared, marketed, and sold [6], [7]. Established tools such as customer relationship management (CRM), social media, and online booking systems are now part of routine operations. Generative AI adds another layer. It can support content creation, customer communication, and personalized recommendations, but adoption depends on organizational readiness, skills, and trust in AI outputs [8], [9], [10].

Sustainability creates a third challenge. Tourism policy increasingly emphasizes resource efficiency, responsible consumption, and local value creation [11], [12]. However, the literature documents an attitude–behavior gap: travelers may support sustainable tourism while still prioritizing price, convenience, and availability [13], [14]. Travel agencies can help narrow this gap by communicating sustainable options and guiding customer choices [15].

These issues are often studied separately. Less is known about how they appear together within tourism distribution. This study therefore asks: (1) How severe are skilled labor shortages among travel distribution organizations? (2) Which digital technologies are adopted, and what role is expected for AI? and (3) How are sustainability practices and customer demand perceived? The paper provides an integrated empirical view based on German industry data.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, 5–9 October 2026, Ljubljana, Slovenia
© 2025 Copyright held by the owner/author(s).
<https://doi.org/10.70314/is.2026.ai4sdg.32>

2 Related work

The tourism workforce literature shows that labor shortages are not simply a short-term recruitment problem. They also reflect retention, career development, and working conditions [4]. [3] emphasizes the continuing importance of human resources in service-intensive tourism. More recent work links workforce recovery to job quality and professional development [4]. The OECD also identifies labor and skills constraints as a continuing issue for tourism businesses [5].

Digital transformation has altered tourism distribution without eliminating intermediaries. Instead, agencies increasingly combine human advice with digital communication and transaction tools [1], [6]. Social media supports customer interaction and visibility, while CRM and booking systems support operational efficiency. Newer technologies, including AI, may extend these capabilities through automation and personalization content [9], [10]. However, smaller firms may face resource and capability constraints [8].

Sustainability presents a different adoption pattern. Basic operational actions, such as reducing paper use, are easier to implement than strategic changes involving staff training, product design, supplier relations, or community engagement. The literature also shows that favorable consumer attitudes do not automatically produce sustainable purchases [13], [14]. Marketing and trusted intermediaries can help by making sustainable choices more visible and understandable [15].

3 Methodology

The study uses a quantitative, cross-sectional design. Data were collected through a standardized online survey between 27 May and 1 July 2024 within the TravelFuture research project. The questionnaire was developed from the tourism management literature and refined with industry experts. It covered workforce shortages, digital technology adoption and perceived value, sustainability practices, customer demand, and organizational characteristics.

The survey was distributed through professional tourism networks, industry associations, and direct invitations to organizations in the German travel distribution sector. Participation was voluntary and anonymous. The final sample included 98 valid responses from traditional travel agencies, tour operators, online travel agencies, and tourism technology providers. Travel agencies formed the largest respondent group.

The analysis used frequencies, percentages, and mean scores. Open-ended comments were examined thematically to complement the descriptive results. The study does not test causal relationships or differences between respondent groups. The sample is non-probabilistic and primarily German, so the findings should be interpreted as an industry snapshot rather than population estimates [16].

4 Results

4.1 Workforce challenges

Skilled labor shortages are a major concern. Sixty-four percent of respondents rated the shortage of qualified personnel as highly relevant. The problem extends beyond recruitment: 97% reported difficulties retaining qualified employees after training, and 83% reported that too few apprentices are entering the profession. Among proposed measures, higher salaries received the highest mean importance score, followed by flexible working conditions and additional benefits.

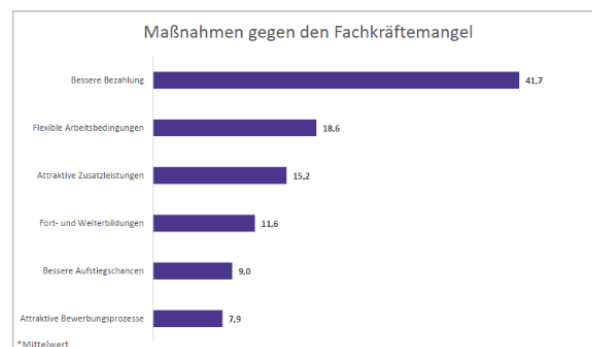
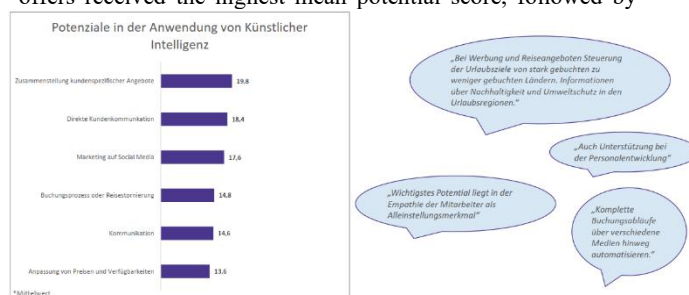


Figure 1: Perceived effectiveness of measures to address skilled labor shortages in travel agencies

4.2 Digitalization and AI

Digital adoption is uneven. CRM systems are used by 81% of respondents, followed by social media (75%) and online booking or management systems (73%). In contrast, chatbots are used by 7% and virtual or augmented reality by 3%. Only 21% reported plans to introduce AI, while 20% planned virtual consultation services and 19% mobile applications. Forty-three percent reported no current use or planned adoption of the emerging technologies included in the survey.

Respondents nevertheless saw value in AI. Personalized travel offers received the highest mean potential score, followed by



direct customer communication and support for social media marketing. The results suggest a gap between perceived potential and organizational readiness.

Figure 2: Perceived benefits of artificial intelligence in travel agencies

4.3 Sustainability practices

Sustainability is present but remains concentrated in basic operational measures. Fifty-nine percent reported largely

paperless administrative processes. Only 37% reported sustainability training and the same share reported offering dedicated sustainable travel products. Thirty percent provide customers with information on environmentally responsible options, while 63% do not participate in community-based or nonprofit sustainability initiatives.

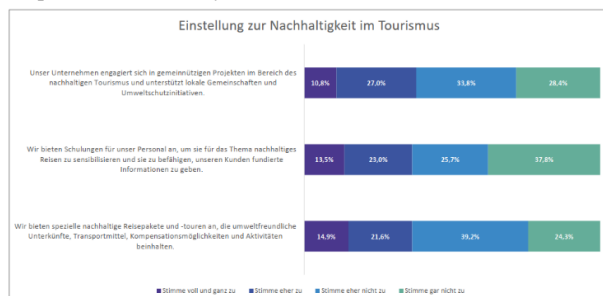


Figure 3: Attitudes toward sustainability in tourism (a)

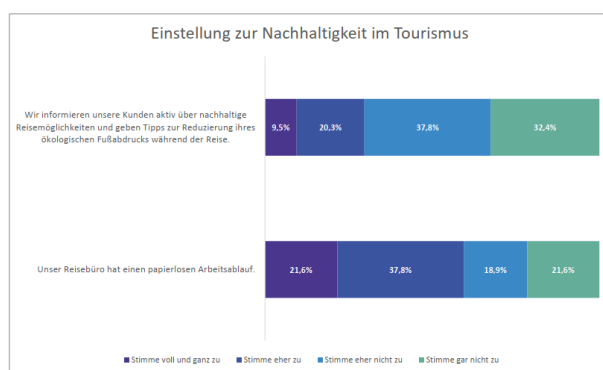


Figure 4: Attitudes toward sustainability in tourism (b)

Perceived customer demand is also weak. Ninety-five percent considered current demand for sustainable travel products below expectations. Land-based alternatives remain underdeveloped: 69% do not develop rail or other ground-based products, 73% do not communicate their environmental benefits, 8% provide related discounts, and 10% organize promotional events. One-third promote land-based travel through social media.

5 Discussion

The findings show that the three challenges are connected. Workforce shortages limit the capacity to implement new technologies and develop new services. Digital tools can reduce routine work and improve customer communication, but advanced technologies require skills, training, and organizational readiness. AI therefore should not be viewed only as a technology purchase. Its value depends on the people and processes around it [9], [10].

The digital results also reveal a clear adoption gap. CRM, social media, and online booking are established, while AI, chatbots, and immersive technologies remain marginal. This pattern is consistent with a gradual transition rather than a sudden technological shift. For many agencies, the next step may be less about adding more tools and more about integrating existing systems with staff capabilities [6], [7], [8].

Sustainability shows a similar implementation gap. Paperless work is relatively common, but practices that require changes in employee behavior, product design, customer communication, or community engagement are less widespread. The weak perceived demand reported by respondents may reinforce this pattern. If customers are seen as unwilling to pay more or change their behavior, agencies have less incentive to invest in sustainable offers. This is consistent with the attitude–behavior gap identified in prior research [13], [14].

The findings also point to a role for agencies as intermediaries of sustainable consumption. They can explain the value of sustainable options, reduce information barriers, and connect travelers with appropriate products [15]. Digital channels can support this role through targeted communication and personalized recommendations. AI may eventually strengthen these functions, but its use should remain transparent and supported by human expertise.

From a management perspective, an integrated approach is preferable to three separate agendas. Workforce policies should improve retention and digital skills. Technology strategies should focus on useful applications rather than adoption for its own sake. Sustainability should move beyond paper reduction toward product development, staff training, customer information, and cooperation with local actors. These capabilities can reinforce each other and contribute to more resilient tourism distribution.

6 Conclusion

This study examined skilled labor shortages, digital transformation, AI, and sustainability in the German travel distribution sector. The evidence shows a sector in transition. Workforce shortages remain serious, established digital tools are widely used, and AI is viewed as promising but is still at an early stage. Sustainability practices are present but tend to focus on basic operational measures, while strategic initiatives and customer-facing actions remain limited.

The main implication is that digital and sustainable transition cannot be separated from workforce development. Agencies need employees who can use new technologies, communicate value to customers, and translate sustainability goals into concrete travel offers. AI can support this process, but adoption should be guided by organizational readiness, staff capabilities, and responsible use.

The study is limited by its non-probability sample, its focus on Germany, its cross-sectional design, and its reliance on self-reported data. Future research should compare countries, follow firms over time, and examine the organizational conditions that enable AI and sustainable business models. Traveler-level evidence would also complement the present supply-side perspective.

Acknowledgments

The authors gratefully acknowledge CENTOURIS, University of Passau, and the ReiseZukunft project partners for their support and for providing access to the survey data used in this study.

References

- [1] R. Law, D. Buhalis, and C. Cobanoglu, “Progress on information and communication technologies in hospitality and tourism,” *International Journal of Contemporary Hospitality Management*, vol. 26, no. 5, pp. 727–750, Jul. 2014, doi: 10.1108/IJCHM-08-2013-0367.
- [2] R. Law, R. Leung, A. Lo, D. Leung, and L. H. N. Fong, “Distribution channel in hospitality and tourism: Revisiting disintermediation from the perspectives of hotels and travel agencies,” *International Journal of Contemporary Hospitality Management*, vol. 27, no. 3, pp. 431–452, Apr. 2015, doi: 10.1108/IJCHM-11-2013-0498.
- [3] T. Baum, “Human resources in tourism: Still waiting for change? – A 2015 reprise,” *Tourism Management*, vol. 50, pp. 204–212, Oct. 2015, doi: 10.1016/j.tourman.2015.02.001.
- [4] T. Baum, S. K. K. Mooney, R. N. S. Robinson, and D. Solnet, “COVID-19’s impact on the hospitality workforce – new crisis or amplification of the norm?,” *International Journal of Contemporary Hospitality Management*, vol. 32, no. 9, pp. 2813–2829, Jul. 2020, doi: 10.1108/IJCHM-04-2020-0314.
- [5] OECD, “OECD Tourism Trends and Policies 2024,” *OECD Tourism Trends and Policies*, vol. 2024, Jul. 2024, doi: 10.1787/80885d8b-en.
- [6] D. Buhalis and R. Law, “Progress in information technology and tourism management: 20 years on and 10 years after the Internet—The state of eTourism research,” *Tourism Management*, vol. 29, no. 4, pp. 609–623, Aug. 2008, doi: 10.1016/j.tourman.2008.01.005.
- [7] M. Sigala, “New technologies in tourism: From multi-disciplinary to anti-disciplinary advances and trajectories,” *Tourism Management Perspectives*, vol. 25, pp. 151–155, Jan. 2018, doi: 10.1016/j.tmp.2017.12.003.
- [8] D. Buhalis, “Technology in tourism-from information communication technologies to eTourism and smart tourism towards ambient intelligence tourism: a perspective article,” *Tourism Review*, vol. 75, no. 1, pp. 267–272, Sep. 2019, doi: 10.1108/TR-06-2019-0258.
- [9] Y. K. Dwivedi *et al.*, “Opinion Paper: ‘So what if ChatGPT wrote it?’ Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy,” *International Journal of Information Management*, vol. 71, p. 102642, Aug. 2023, doi: 10.1016/j.ijinfomgt.2023.102642.
- [10] V. Venkatesh, J. Thong, and X. Xu, “Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology,” *Management Information Systems Quarterly*, vol. 36, no. 1, pp. 157–178, Mar. 2012.
- [11] UN Tourism, “UN Tourism Tourism Highlights.” Accessed: Jul. 29, 2026. [Online]. Available: <https://www.untourism.int/tourism-statistics/tourism-statistics-database>
- [12] WEF, “The promise and challenges of sustainable travel,” World Economic Forum. Accessed: Jul. 29, 2026. [Online]. Available: <https://www.weforum.org/stories/industries-in-depth/sustainable-travel-and-tourism/>
- [13] A. Budeanu, “Sustainable tourist behaviour – a discussion of opportunities for change,” *International Journal of Consumer Studies*, vol. 31, no. 5, pp. 499–508, 2007, doi: 10.1111/j.1470-6431.2007.00606.x.
- [14] E. Juvan and S. Dolnicar, “The attitude–behaviour gap in sustainable tourism,” *Annals of Tourism Research*, vol. 48, pp. 76–95, Sep. 2014, doi: 10.1016/j.annals.2014.05.012.
- [15] X. Font and S. McCabe, “Sustainability and marketing in tourism: its contexts, paradoxes, approaches, challenges and potential,” *Journal of Sustainable Tourism*, vol. 25, no. 7, pp. 869–883, Jul. 2017, doi: 10.1080/09669582.2017.1301721.
- [16] ReiseZukunft, “Reisebürobefragung 2024: Fachkräftemangel, neue Technologien & Nachhaltigkeit [Travel agency survey report].” Accessed: Jul. 29, 2026. [Online]. Available: <https://reisezukunft.de/forschungsergebnisse/reisebueroebefragung/>

Indeks avtorjev / Author index

Achuthan Krishnashree	13
Age Terungwa James	21
Ali Abbas Syed Muhammad	25
Annamali Devisakti	44
Badiger Prakash K.	40
Bansal Sneh	40
Benjelloun Andaloussi Omar	87
Bogoeski Teodor	64
Bong Yang	44
Bonidia Robson Parmezan	48
Bosco Francesca	52
Carvalho André Carlos Ponce de Leon Ferreira	48
Cellérier Leïla	44
Chen Minjia	44
Cruz Wellington	48
Erokhin Dmitry	56
Ezzatian Shohreh	79
Florentino Bruno Rafael	48
Fuse Tomoyuki	60
Gams Matjaž	25
Geoffre Thierry	36
Geoffre Trystan	36
Ghedhioui Imen	17
Gulati Asheesh	36
Guruvayur Sivaramakrishnan R	83
Hallensleben Sebastian	52
Idowu Peter Damilare	72
Iqbal Muhammad Zahid	44
Jankulovski Hristijan	64
Jaris Julieyana	29
Jogan Sushma N.	40
Kani Kolil Vysakh	13
Kobayashi Masamichi	60
Komendantova Nadejda	56
Luyao Zhang	9
Maier Janine	87
Mang Stefan	87
Marhraoui Mohamed Amine	87
Martins João	17
Meyerhoff Nielsen Morten	17
Mohd Yusuf Mohammad Rizal	29
Nasteski Zoran	64
Nyalala Innocent	5
Oda Kazue	60
Oda Masato	60
Parvathy S. U.	13
Peña-Ramírez Camilo	87
Pitogo Vicente	44
Quintiliano Leonardo David	68
R. Guruvayur Sivaramakrishnan	13
Rajeev Devi	83
Saha Rohan	5
Sailaja Neelima	44
Sanoussi Bendjedid Rachad	87
Sestito Camila	48

Shaikh Iqra	17
Tada Jin	44
Taya Sho.....	60
Wang Yutian	9
Yeo Bernice.....	33
Yunus Sharina	75