

Youngsam Chun

Responsible AI Researcher · AI Safety, Governance, and Value Alignment

✉ deep1003@snu.ac.kr · ys.chun@kt.com ☎ (+82) 010-3458-0070

✉ Frontier AI Lab, KT, Taebong-ro 151, Seocho-gu, Seoul 06763, Republic of Korea

ORCID: [0000-0002-6877-6230](https://orcid.org/0000-0002-6877-6230) · [Google Scholar](#) · [GitHub: deep1003](#) · [LinkedIn](#)

PROFILE

AI safety and governance researcher with applied industry experience at KT Corporation. My work spans red-teaming and defence of retrieval-augmented generation systems, responsible AI governance and risk taxonomies for high-risk and Physical AI systems, and value alignment and evaluation of large language model judges. I combine rigorous evaluation methods (adversarial testing, alignment auditing and structured risk taxonomy design) with policy-facing governance frameworks, and I aim to advance AI safety, fairness and governance research within the UK's AI governance research community.

RESEARCH AREAS

- AI safety, red-teaming, and adversarial evaluation of LLM and RAG systems
- Responsible AI governance and risk taxonomies for high-risk and Physical AI systems
- Value alignment and evaluation of LLM judges (moral orientation and calibration)
- Agentic RAG security, black-box poisoning attacks, and defences
- Physical AI risk, embodied AI, and cyber-physical system safety
- Patent analytics, scientometrics, and AI innovation measurement
- Network science, economic complexity, and knowledge-space modelling

EDUCATION AND TRAINING

PhD, Technology Management, Economics, and Policy (TEMEP)

Mar 2019 – Aug 2024

Seoul National University, Seoul, Korea

Final grade 4.28/4.5 (97.8); EQF 8

Thesis: *Exploring complementary innovation of AI technology from a multidimensional knowledge network approach* (HAL: tel-05047941)

MSc, Technology Management, Economics, and Policy (TEMEP)

Mar 2017 – Feb 2019

Seoul National University, Seoul, Korea

Final grade 4.33/4.5 (98.3); EQF 7

Thesis: *Heterophily Effects on Industrial Innovation Leadership Changes in Autonomous Vehicle Industry*

BSc, Telecommunication and Information Engineering

Mar 1990 – Feb 1997

Korea Aerospace University, Gyeonggi-do, Korea

Final grade 3.44/4.5 (76.4); EQF 6

Signal processing, wireless communication and network protocols; data transmission, information security and digital communication; avionics and satellite communications.

Economic Networks Summer School

Jun 2024

University of Oxford, Oxford, United Kingdom · complex network analysis, economic complexity and geography.

Masterclass: MSCA Post-Doctoral Fellowships

May 2025

Université Paris 1 Panthéon-Sorbonne, Paris, France · intensive proposal-writing workshop.

WORK EXPERIENCE

Responsible AI Researcher · Responsible AI Department, Frontier

Aug 2024 – Present

AI Lab, KT Corporation

Seoul, Korea

- Establishment of a responsible AI governance framework for high-risk domains.
- Joint research and development of a Korea-specific custom GPT-4o with Microsoft.
- Development of value-alignment technology reflecting Korea's cultural context.

Adjunct Professor in Business Analytics · UNIST

Aug 2024 – Present

Graduate School of Technology and Innovation Management, Ulsan National Institute of Science and Technology, Ulsan, Korea

- Graduate teaching on business analytics and machine/deep learning using Python.
- Supervision of graduate students in thesis writing and research.

SELECTED PROJECTS

Physical AI Risk Taxonomy

2026 – Present

Project Lead, Frontier AI Lab, KT (Korea). A public, bilingual evidence map of risk cards for AI systems that perceive, decide and act in the physical world, from robots and humanoids to drones and autonomous vehicles. Built with a hybrid method that pairs a human-defined risk hierarchy with bottom-up discovery from a large corpus of papers, patents and policy documents, and released with alignment tags, audit logs and a public technical report. [Live site](#) · [Repository](#)

Physical AI Risk Assessment for Urban Deployment · with MIT

Senseable City Lab

Aug 2026 – Present

Project Lead, Frontier AI Lab, KT (Korea), with the MIT Senseable City Lab and the City of Seoul. A twelve-month project that applies the Physical AI Risk Taxonomy to real urban settings, building a risk-assessment framework for how embodied AI and street-level conditions interact in a city, and working towards a Korea-specific evaluation standard.

Red-Teaming and Defence of Retrieval-Augmented Generation (RAG) Systems

Oct 2025 – Nov 2026

Project lead at Frontier AI Lab, KT (Responsible AI Department), with Korea University (Prof. Buru Chang's group). Black-box red-teaming of agentic RAG systems (the KidnapRAG attack) and corresponding defences, delivering the ARS-Bench security benchmark, an open-source toolkit and a registered patent, and connecting adversarial security research to the responsible deployment of B2B RAG services.

Contextual Value Alignment through Genre-Based Aesthetic

Judgment

2026 – Present

International collaboration with Isa Weisha Wang (Institute for Ethics in AI, Technical University of Munich). A framework for measuring where large language model evaluators diverge from human interpretive communities in aesthetic judgment, using film genres as measurable cultural contexts. It distinguishes universal from context-dependent value dimensions and locates human and AI gaps through a sign-flip test and a subspace-resolved orientation and calibration audit. Companion to the orientation and calibration work, with a joint manuscript in preparation.

AI Topic Space: Policy, Science and Technology in the Korean AI Knowledge Ecosystem

2025 – Present

Principal investigator, with STEPI and KDI (Korea). An interactive semantic map of AI topics across policy reports, papers and patents that lets users explore how the science, technology and policy agendas align and where Korean policy falls behind. [Live site](#) · [Repository](#) · [HuggingFace dataset](#)

Innovation Characteristics of Generative AI and Large Language

Models: A Patent Analytics Approach

Jul 2025 – Jan 2026

Research Collaborator, KT Frontier AI Lab (Responsible AI Department), in collaboration with STEPI. Sentence-BERT embeddings (PatentSBERTa) and ensemble modelling (XGBoost + MLP) with technological-distance clustering to assess dominant design emergence and cross-national competitive dynamics.

Analysis of Industrial AI Innovation and National Competitiveness: A Scientometric Approach

Dec 2025 – Jan 2026

Research Collaborator, Seoul National University & Korea Development Institute (KDI). Patent analytics and the Economic Complexity Index (RCA-based) to study domain specialisation and provide policy frameworks for trustworthy AI governance and industrial co-evolution.

PUBLICATIONS AND PREPRINTS

KidnapRAG: A Black-Box Attack for Hijacking Reasoning in Agentic Retrieval-Augmented Generation Systems	2026
Choi, C., Kim, E., Lee, K., Chun, Y. , Jeong, J., Kim, E., Oh, M., Jang, J., & Chang, B. (2026). Industry-academia collaboration with Prof. Buru Chang's group (Korea University); Chun, Y. led the KT side of the project and holds first-inventor priority on the registered Korean patent (Jul 2026). <i>Under review at EMNLP 2026</i> . arXiv:2607.00422 · Code	
ARS-Bench: Benchmarking Security Risks Across the Agentic RAG Attack Lifecycle	2026
Choi, C., Kim, E., Lee, K., Chun, Y. , Jeong, J., Kim, E., Oh, M., Jang, J., & Chang, B. (2026). <i>ACL ARR August 2026 submission</i> . A benchmark spanning the full agentic RAG attack lifecycle, from adversarial document ingestion to final response generation, structured along five security-critical dimensions. Paper	
Silent Failure of Moral Alignment in Large Language-Model Evaluators	2026
Chun, Y. <i>Under review at Scientific Reports</i> (June 2026). Reproducibility release: deep1003/silent-failure-moral-alignment .	
Moral Orientation and Calibration: Coupled in Human Annotators, Separable in Judge LLMs	2026
Chun, Y. (2026). Moral Orientation and Calibration: Coupled in Human Annotators, Separable in Judge LLMs. <i>ICML 2026 Pluralistic Alignment Workshop</i> (accepted). Paper · Code	
When Fear Becomes Beauty: Genre as Context in Human and AI Aesthetic Judgment	2026
Wang, I. W., & Chun, Y. (2026). Manuscript in preparation. A technical framework for contextual value alignment of large language model evaluators in the aesthetic domain, extending the orientation and calibration decomposition to genre-typed cultural contexts. Cross-institutional collaboration between the Institute for Ethics in AI (TU Munich) and KT Frontier AI Lab.	
A Hybrid Corpus-to-Taxonomy Pipeline for Physical AI Risk Cards	2026
Chun, Y. (2026). Technical Report v1.0, Responsible AI Department, Frontier AI Lab, KT Corporation. Documents the data collection, deduplication and hierarchical assignment pipeline behind the Physical AI Risk Taxonomy, with reproducibility artefacts.	
AI Technology Specialization and National Competitiveness	2024
Chun, Y. , Hur, J., & Hwang, J. (2024). <i>PLOS ONE</i> . Link	
The Nexus of Artificial Intelligence and Green Innovation: A Cross-Density Analysis at the Country Level	2024
Chun, Y. , & Hwang, J. (2024). <i>Journal of the Knowledge Economy</i> . Link	
The Moderating Role of Design Innovation in the Relationship Between Technology Complexity and Firm Performance	2024
Hur, J., Chun, Y. , Hwang, J., & Kim, K. (2024). <i>Technology Analysis & Strategic Management</i> , 1–14. Link	
PATENTS	
Method and Apparatus for Automatically Constructing a Physical AI Risk Taxonomy	Filed Aug 2026
Chun, Y. et al. (2026) [P20260372]. Korean patent application (filed 3 Aug 2026) covering the automated construction of a risk classification system for Physical AI, arising from the Physical AI Risk Taxonomy project.	
Method, Server and Computer Program for Dynamically Adjusting Value Information	2025 – Present
Chun, Y. , Park, Y., Yoon, J., & Lee, S. (2025) [P20250217]. A value-alignment method for generative AI combining multi-agent debate and reinforcement-learning-based reward modelling to align dynamically with evolving ethical and social values without continuous human intervention.	

KidnapRAG: Black-Box Attack for Hijacking Reasoning in Agentic

RAG Systems

Registered Jul 2026

Chun, Y. (first inventor), Choi, C., Kim, E., Lee, K., Jeong, J., Kim, E., Oh, M., Jang, J., & Chang, B. (2026). Registered Korean patent (Jul 2026) on black-box poisoning attacks against agentic retrieval-augmented generation systems, with Chun as first inventor.

HONOURS AND AWARDS

- **Travel Grant Winner**, The Responsible AI Forum (TRAIF), Technical University of Munich · 2026.
- **Excellent Patent Award**, Invention Competition 2025, KT Corporation R&D Center · Aug 2025 (patent *Method, Server and Computer Program for Dynamically Adjusting Value Information*).
- **Best PhD Paper Award**, KOSIME 2024, Korea Society of Innovation, Management, and Economics · Jul 2024.

CONFERENCES AND SEMINARS

- **OECD, Paris** (Oct 2026), with the KDI delegation · meeting the STI AI and Emerging Digital Technologies Division (1 Oct) and presenting at the KDI-OECD Expert Roundtable on AI, Innovation and Productivity (2 Oct), on multi-dimensional AI risk classification.
- **Korean Association for Policy Studies (KAPS) 2026 Summer Conference**, Korea (14 Aug 2026) · “Policy-Gap Visualisation,” STEPI session on inclusiveness and policy-agenda gaps in the AI transition, within the NRC national AI collaborative research programme.
- **NRC Outcomes Forum on the AI Transition of Industry and Social Institutions**, Seoul (7 Oct 2026) · “Discourse Analysis of the State and Gaps in AI Debate,” invited talk at the NRC national AI collaborative research programme dissemination forum.
- **ICML 2026 Pluralistic Alignment Workshop** (Poster Session), 2026 · “Moral Orientation and Calibration: Coupled in Human Annotators, Separable in Judge LLMs.”
- **2024 Seminar on Artificial Intelligence in Budapest**, Hungary (Oct 2024) · “Strategic Approaches to AI Development and Implementation.”
- **KOSIME 2024**, Jeju, Korea (Jul 2024) · “National AI Capabilities and Product Diversification: A Multidimensional Knowledge Space Approach.”
- **6th Geography of Innovation Conference (GEOINNO 2022)**, Milan, Italy (Jul 2022) · “Economic Resilience of EU Regions to Localized Attacks.”
- **ACM SIGCHI 2020 (Fair & Responsible AI Workshop)**, Hawaii (Apr 2020) · “How Does Perceived Fairness Differ? Exploring Multicultural Landscape for Ethical AI Design.”
- **The Future Use of Autonomous Vehicles** (Invited Seminar), Paris/EHESS (Jun 2019) · “Technology Shock, Microdynamics, and Industrial Leadership Change.”

PROFESSIONAL SERVICE

Journal Reviewer

2023 – Present

Scientific Reports, PLOS ONE, Journal of the Knowledge Economy, European Research on Management and Business Economics, Business Strategy and the Environment, Research in Globalization, and Humanities and Social Sciences Communications.

AI Governance & Innovation Policy Consulting

2024 – Present

Invited Speaker, KDI (Nov 2024); AI Industry Roundtable, Seoul AI Hub (Aug 2024); AI patent/paper data analysis, KDI (Oct 2024); AI regulation research, STEPI (Jun 2024); AI-industry patent analysis, STEPI (Apr 2024).

Professional Affiliations & Networking

2024 – Present

GSMA AI for Impact Taskforce · Responsible AI Maturity Roadmap; standardised benchmarks across five key dimensions.

Policy and Industry Advisory

2023 – Present

Expert Advisor, HYGINO (carbon-footprint policy design); Advisor, National Grand Challenges Energy Division, STEPI; Advisory Committee & Author, KOTRA Global Economic and Trade Outlook Report.

TEACHING EXPERIENCE

- **Lecturer · Business Analytics**, Graduate School of Technology and Innovation Management, UNIST (Aug 2024 – Present): hands-on data analysis in Python/R, AI and generative-AI foundations, group-project supervision.
- **Invited Lecturer · Generative and Agentic AI: Innovation Mechanisms and Ethics**, Pukyong National University (Mar 2026): the evolutionary trajectory of AI from perception to generative, agentic, and physical AI; agentic AI mechanisms (multimodal interaction, reasoning/planning, action execution, self-learning); and the ethics and regulatory landscape of generative AI.
- **Invited Lecturer · AI Co-evolution and Innovation Dynamics in Mobile Technology**, Underwood International College, Yonsei University (Nov 2025): evolution of mobile/communication technologies from 1G to 5G/6G, energy and security challenges.
- **Invited Lecturer · Generative and Agentic AI: Innovation Mechanisms and Ethics**, Kyung Hee University (Mar 2025).

SKILLS

Programming: Python (ML, data science, NLP), SQL, R, LaTeX.

AI & NLP: Sentence-BERT, PatentSBERTa, RAG systems.

Analytics: network analysis, embeddings, clustering, UMAP, scientometrics, patent analytics.

Statistics & visualisation: STATA, Tableau.

LANGUAGE SKILLS

Korean · native. **English** · C1/C2 (Listening C2, Reading C2, Spoken C1, Writing C2). **Japanese** · B2/C1.

REFERENCES

- **Prof. Chris Russell** · Dieter Schwarz Associate Professor in AI, Government and Policy, University of Oxford (Oxford Internet Institute). chris.russell@oii.ox.ac.uk
- **Prof. Christoph Lütge** · Professor of Business Ethics; Director, Institute for Ethics in AI (IEAI), Technical University of Munich. christoph.luetge@tum.de
- **Prof. Junseok Hwang** · Professor, TEMEP, Seoul National University; MSc and PhD supervisor. junhwang@snu.ac.kr
- **Dr Jongrip Kim** · Team Leader, Research Impact and Dissemination Team, Science and Technology Policy Institute (STEPi); principal investigator of the NRC national AI collaborative research programme. lipk@stepi.re.kr
- **Prof. Keungoui Kim** · Professor, Underwood International College, Yonsei University. awekim@yonsei.ac.kr
- **Prof. Camille Salinesi** · University Professor, Université Paris 1 Panthéon-Sorbonne; Co-Director, Observatoire de l'IA. camille.salinesi@univ-paris1.fr