

Jaehong Kim

Ph.D. Candidate at KAIST (Expected Graduation: February 2027)
luke.4.18@kaist.ac.kr | <https://jaehong-k.github.io/>

RESEARCH INTERESTS

I study how moral emotions shape **human and machine behavior**. I build **multimodal frameworks** that detect moral emotions in English and Korean text and images, and find that **moral outrage** (anger at perceived wrongdoing) most strongly predicts committed user participation and is disproportionately used by suspected **foreign-linked trolls**. In **large language models (LLMs)**, I show that steering internal representations of moral outrage causally shifts punishment judgments, leading models to recommend harsher sentences for identical offenses.

Beyond academia, my research on foreign-linked influence operations has received **national media coverage**, including an [SBS News](#) report with over 1.2 million views, and has **contributed to policy discussions** among lawmakers.

EDUCATION

Korea Advanced Institute of Science and Technology (KAIST) <i>Ph.D., Graduate School of Culture Technology, Social Computing Lab (Advisor: Wonjae Lee)</i> Dissertation: How Moral Outrage Shapes Human and Machine Behavior	Daejeon, Korea 2023.08 - Present
<i>M.S., Graduate School of Culture Technology, Social Computing Lab (Advisor: Wonjae Lee)</i> Thesis: How moral emotions shape the commitment of political support in the online petition	2021.08 - 2023.08
Handong Global University <i>B.A., Communication, Psychology, ICT Convergence (Minor), Cum Laude</i>	Pohang, Korea 2015.03 - 2021.08

SELECTED PUBLICATIONS

Full list: [\[Google Scholar\]](#)

Peer-Reviewed Publications (* denotes co-first authorship, † denotes international collaboration)

- **J. Kim***, H. Kim*, J. Kim, A. Oh, T. Holz, W. Lee, M. Cha. (2026). Cross-National Information Attacks: A Two-Decade Analysis of Troll Behavior in Korea. *In Proc. of the 35th USENIX Security Symposium (USENIX Security '26)*. [\[Zenodo\]](#) †
- S. Park*, **J. Kim***, H. Kim, H. Bin, S. Moon, W. Lee. (2026). Moral Outrage Shapes Commitments Beyond Attention: Multimodal Moral Emotions on YouTube in Korea and the US. *In Proc. of The Web Conference (WWW 2026)*. [\[GitHub\]](#) †
- M. Sundriyal*, **J. Kim***, W. Dong, W. Lee, M. Cha. (2026). Nothing to Dread: A Measurement Study of a Dark Web Social Forum. *In Proc. of the 2026 ACM Internet Measurement Conference (IMC '26)*. [\[Zenodo\]](#) †
- W. Dong*, M. Sundriyal*, S. Park*, **J. Kim***, M. Cha, T. Chakraborty, W. Lee. (2025). Parallel Communities Across the Surface Web and the Dark Web. *In Findings of Empirical Methods in Natural Language Processing (EMNLP Findings 2025)*. *Nominated for Best Resource Paper*. [\[Zenodo\]](#) †
- **J. Kim***, C. Jeong*, S. Park*, M. Cha, W. Lee. (2024). How Do Moral Emotions Shape Political Participation? A Cross-Cultural Analysis of Online Petitions Using Language Models. *In Findings of the Association for Computational Linguistics (ACL Findings 2024)*. [\[GitHub\]](#) †
- J. Kim, J. Kwon, L. F. Vecchietti, W. Dong, **J. Kim**, M. Cha. (2026). Machine Behavior in Relational Moral Dilemmas: Moral Rightness, Predicted Human Behavior, and Model Decisions. *In Findings of the Association for Computational Linguistics (ACL Findings 2026)*. [\[GitHub\]](#) †

Under Review & Working Papers

- **J. Kim**, J. Kim, H. Kim, J. Kwon, M. Cha, W. Lee. (2026). Investigating the Causal Role of Internal Moral Outrage in LLM Punishment Judgments. *Under review at ICLR 2027*. †
- **J. Kim**, S. Park, W. Dong, A. Zhunis, M. Cha, W. Lee. From Online Outrage to Offline Protest: Evidence from the Israel-Palestine and Russia-Ukraine Conflicts. *Working paper*. †

RESEARCH EXPERIENCE

Max Planck Institute for Security and Privacy (MPI-SP)	Bochum, Germany
<i>Research Intern, Data Science for Humanity (Advisor: Meeyoung Cha)</i>	2025.09 - 2026.02
<i>Research Intern, Data Science for Humanity (Advisor: Meeyoung Cha)</i>	2024.09 - 2024.11

INTERNATIONAL COLLABORATIONS

Max Planck Institute for Security and Privacy (MPI-SP)	2024-Present
<i>Data Science for Humanity Group (Prof. Meeyoung Cha); research on influence operations and dark web social media analysis</i>	
Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)	2025-Present
<i>Department of NLP (Prof. Alham Fikri Aji); explainable classification of moral emotions</i>	

INVITED TALKS

NYU Center for Conflict and Cooperation	2026
Max Planck Institute for Security and Privacy, Data Science for Humanity Group	2026
Reichman University, AI/NLP Seminar	2026
Korea Computer Congress Top Conference Session	2026
Max Planck Institute for Security and Privacy, Workshop on the Societal Impact of AI Systems	2025

FELLOWSHIPS & AWARDS

KAIST Fake News Challenge Best Prize (KRW 2,000,000)	2026
Global Advanced Human Resource Development Program in Information Security, IITP Competitive national fellowship; funded research visit to MPI-SP (KRW 18,000,000)	2025
Hyundai Motor Chung Mong-Koo Global Scholarship (KRW 28,000,000)	2025
Adventure Research Program, KAIST (KRW 1,000,000)	2024
Young-Han Kim Global Leader Scholarship, KAIST (KRW 4,000,000)	2023

TEACHING ASSISTANT

Natural Language Processing for the Humanities and Social Sciences (HSS510)	2025.03 – 2025.06
<i>Led lab sessions on LLM fine-tuning, covering core concepts and hands-on implementation</i>	
Social Network Analysis (CTP471)	2024.03 – 2024.06
<i>Assisted student projects on network theory and analysis in R, providing feedback and technical support</i>	
Social Computing (GCT576)	2023.09 – 2023.12
<i>Mentored student research applying core sociological concepts to computational social science</i>	
Computational Social Science (HSS312), Selected as Best TA	2021.09 – 2021.12
<i>Led coding tutorials on data collection methods (e.g., web scraping, APIs)</i>	

ACADEMIC SERVICE

Panelist: Global South ML Workshop at ICML 2026 / **Reviewer:** ICLR 2027, CSCW 2025 / **Shared Task Organizer:** FinNLP-FNP-LLMFinLegal 2025