Jin Hwa Lee

PhD student in the intersection of Cognitive Neuroscience & Machine Learning
University College London

Sainsbury Wellcome Centre & Gatsby Computational Neuroscience Unit, 25 Howland St. London W1T 4JG

✉ jin.lee.22@ucl.ac.uk | ⌂ jinhl9@github.io

Research Statement

My research focus on fundamental understanding of neural network model-how the structure of data and the inductive biases of models shape learning. Particularly, how compositional world models might emerge from this interplay. My work blends theory, like tools from statistical physics, with empirical studies of large language models.

Recently, I've been interested in more pressing problem: how can we steer increasingly capable AI systems toward outcomes we actually want? I'm seeking mentorship and community in AI alignment research, driven by a belief that it's a critical challenge for all of the citizens and it is where my curiosity and skills can make a meaningful contribution.

Education

University College London   *London*
PhD Student   *Oct 2022 - present*
- Supervisor: Prof. Andrew Saxe

Technical University of Munich   *Munich*
MSc Neuroengineering   *Oct 2019 - Feb 2022*
- Supervisor: Prof. Mackenzie Mathis
- Thesis: CEBRA: Multi-Modal Unsupervised Learning of Consistent Embeddings in Neural and Behavioral Activity

Korea Institute of Advanced Science and Technology (KAIST)   *Daejeon*
BSc Physics   *Mar 2015 - Sep 2019*
- Magna Cum Laude

Research Projects

Lee, J. H., Singh, A., Lampinen, A., & Saxe, A. Distinct Computations Emerge From Compositional Curricula in In-Context Learning, *ICLR 2025 Spurious Correlation and Shortcut Learning Workshop*

Lee, J. H.*, Jiralerspong, T*., Yu, L., Bengio, Y., & Cheng, E. (2024). Geometric Signatures of Compositionality Across a Language Model's Lifetime. *Under review*

Dorrell, W., Hsu, K., Hollingsworth, L., **Lee, J. H.**, Wu, Jiajun., Finn, Chelsea., Latham, PE., Behrens, TEJ., & Whittington, JCR.(2025). Range, not Independence, Drives Modularity in Biological Inspired Representation, *ICLR 2025*.

Lee, J. H., Mannelli, S. S., & Saxe, A. (2024). Why Do Animals Need Shaping? A Theory of Task Composition and Curriculum Learning. *ICML 2024*

Schneider, S.*, **Lee, J. H.***, & Mathis, M. W. (2023) Learnable latent embeddings for joint behavioral and neural analysis. *Nature*, 617(7960), 360-368

Servadei, L., **Lee, J. H.**, Medina, J. A. A., Werner, M., Hochreiter, S., Ecker, W., & Wille, R. (2022). Deep reinforcement learning for optimization at early design stages. IEEE Design & Test, 40(1), 43-51.

Mann, K.S., Schneider, S., Chiappa, A., **Lee, J. H.**, Bethge, M., Mathis, A., Mathis, M. W. (2021). Out-of-distribution generalization of internal models is correlated with reward. In Self-Supervision for Reinforcement Learning Workshop-ICLR (Vol. 2021).

Awards, Fellowships, & Grants

2024	**Brain, Minds and Machines 2024 Summer School Travel Grant & Scholarship**, CBMM	*$ 3000*
2024	**COSYNE 2024 Travel Grant**, COSYNE	*$ 1000*
2020	**IEEE Brain BCI Hackathon**, IEEE	*1st Prize*
2020	**DAAD Scholarship**, DAAD	*$ 13,000*
2017	**National Science and Engineering Undergraduate Scholarship**, KOSAF	*$ 11,000*

Invited Talks

Lee, J.H. *2025*. Compositional Learning, COSYNE Workshop, Montreal, Canada

Lee, J.H. *2025* Invited Talk (Host:Emily Cheng & Marco Baroni), Computational Linguistics and Linguistics Theory Group at Universitat Pompeu Fabra, Barcelona, Spain

Itay Evron, **Lee, J.H**. 2024. 3rd Conference on Lifelong Learning Agents (CoLLAs) Tutorial: Theoretical Advances in Continual Learning. Pisa, Italy

Lee, J.H. *2024*. A Tutorial on CEBRA. COSYNE Workshop Talk: Sharpening Our Sight. Cascais, Portugal

Lee, J.H. *2024*. Invited talk, Neuronaut, London.

Lee, J.H *2023*. *Analytical Model of Compositional Learning*. Invited talk, Tim Behrens Lab, SWC, London.

Teaching Experience

Fall 2023	**Systems Neuroscience & Theoretical Neuroscience**, Teaching Assistant	*SWC&GCNU, London*
Summer 2020	**Machine Learning: Methods and Tools**, Teaching Assistant	*TUM, Munich*

Outreach & Professional Development

2024	**Brains, Minds and Machines Summer School, MIT** , Participant	*Woods Hole*
2023-2024	**WiML Mentoring** , Mentor	*Remote*
2023	**Annalytical Connectionism**, Participant	*London*
2016-2019	**Connect Foundation**, Education Volunteer	*Seoul*