Jin Hwa Lee

PhD student @ Theory of Learning Lab, University College London

jin.lee.22@ucl.ac.uk jinhl9.github.io | google scholar

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

Research Statement

My research focuses on understanding how certain structure of data and inductive biases of models shape learning. Particularly, how compositional world models might emerge from this interplay. My work blends theory based on a tractable toy model and tools from statistical physics, with empirical studies of large language models. With such approches, my current projects aim to understand reasoning capabilities of LLMs through a lens of compositionality.

Recently, I've been more committed to be involved in the pressing problems revolving AI capability and alignment. I'm seeking mentorship and community in AI safety research, driven by thoughts that it is a critical challenge that requires the action now and it is where my interest and skills can make a meaningful contribution.

Education

- **University College London** *Oct 2022 - Present*
  PhD student London, UK
  - Supervisor: Prof. Andrew Saxe
- **Technical University of Munich** *Oct 2019 - Feb 2022*
  MSc Neuroengineering Munich, Germany
  - Supervisor: Prof. Mackenzie Mathis
  - Thesis: CEBRA: Multi-Modal Unsupervised Learning of Consistent Embeddings for Neural and Behavioral Activity
- **Korea Advanced Institute of Science and Technology(KAIST)** *Mar 2015 - Sep 2019*
  BSc Physics Daejeon, South Korea
  - Magna Cum Laude

Research

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

- Demonstration of how curricula-like data structure in-context during pretraining can influence models' solution strategy on compositional tasks

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

- Analyzing geometric properties of hidden representations in LLMs throughout pretraining and how does compositional structure of language is reflected and correlated to the linguistic capability

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

- Deriving necessary and sufficient conditions on sample data statics to gain modular representation with biological neural constraints

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

- Analytical study of deterministic policy learning dynamics of compositional RL in high-dimensional teacher-student setup

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

- Contrastive learning and identifiability in ICA inspired multi-modal ML method for mapping high dimensional neural and behavioral data

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

- Solving combinatorial optimization problem using pointer network model and reinforcement learning

Invited Talks

- **COSYNE 2025 Workshop: Compositional Learning** — *Apr 2025*
  Analytical Approach to Study Compositional Learning — Montreal, Canada
- **Invited talk: Learning Dynamics of Linguistic Compositionality** — *Feb 2025*
  Computational Linguistics Group, Universitat Pompeu Fabra, hosted by Marco Baroni & Emily Cheng — Barcelona, Spain
- **3rd Conference on Lifelong Learning Agents (CoLLAs)** — *Jul 2024*
  Tutorial: Theoretical Advances in Continual Learning, ***Itay Evron, Jin Hwa Lee*** — Pisa, Italy
- **COSYNE 2024 Workshop: Sharpening Our Sight** — *Mar 2024*
  CEBRA Tutorial — Cascais, Portugal
- **Invited talk: Tim Behrens group @ UCL, Oxford** — *May 2023*
  Analytical Model of Compositional Learning — London, UK

Awards and Scholarships

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

Teaching Experience

- **Systems Neuroscience & Theoretical Neuroscience** — *Fall 2023*
  Sainsbury Wellcome Centre&Gatsby Computational Neurosicence Unit, Teaching Assistant — London, UK
- **Machine Learning: Methods and Tools** — *Summer 2020*
  Technical University of Munich, Teaching Assistant — Munich, Germany

Outreach & Professional Development

- **Brains, Minds and Machines Summer School** — *Summer 2024*
  MIT CBMM, Participant — Woods Hole, US
- **Women in Machine Learning Mentoring** — *2023-2024*
  Mentor — Remote
- **Analytical Connectionism** — *Summer 2023*
  Participant — London, UK
- **Connect Foundation** — *2016-2019*
  Education Volunteer — Seoul, South Korea