Curriculum Vitae/Resume

Ekdeep Singh Lubana
Email: ekdeeplubana@fas.harvard.edu

Education

Postdoctoral Fellow, Harvard University *August, 2024–ongoing*
(Part of CBS-NTT Program on Physics of Intelligence)
Advisors: Hidenori Tanaka and David Krueger

Ph.D. Candidate, University of Michigan, Ann Arbor *August, 2019–August, 2024*
(Co-affiliated with Center for Brain Science, Harvard University)
Advisors: Robert P. Dick and Hidenori Tanaka

B.Tech., Indian Institute of Technology, Roorkee *July, 2015–May, 2019*
Major: Electronics and Communication Engineering
Thesis: Resource Efficient Techniques for Embedded Machine Vision (Nominated for Best Bachelor's Thesis)

Areas of Interest

· AI Alignment, Science of Deep Learning, Interpretability

Experience

· **Research Intern**, Qualcomm AI Research, Amsterdam *June., 2023–Nov., 2023*
Mentors: Taco Cohen, Johann Brehmer, and Pim de Haan

· **Research Affiliate**, Krueger AI Safety Lab *Aug, 2022–Present*
Mentor: David Krueger

· **Research Intern**, Bell Labs Cambridge, UK *Sept., 2021–Dec., 2021*
Mentor: Akhil Mathur

· **Research Intern**, Physics and Informatics Lab, NTT Research Inc. *May, 2021–Aug., 2021*
Mentor: Hidenori Tanaka

Publications (* denotes equal contribution)

1. Sai Sumedh R. Hindupur*, **Ekdeep Singh Lubana***, Thomas Fel*, and Demba Ba. Projecting assumptions: The duality between sparse autoencoders and concept geometry. *arXiv preprint arXiv:2503.01822*, 2025

2. Thomas Fel*, **Ekdeep Singh Lubana***, Jacob S. Prince, Matthew Kowal, Victor Boutin, Isabel Papadimitriou, Binxu Wang, Martin Wattenberg, Demba Ba, and Talia Konkle. Archetypal sae: Adaptive and stable dictionary learning for concept extraction in large vision models. *arXiv preprint arXiv:2502.12892*, 2025

3. Core Francisco Park*, Andrew Lee*, **Ekdeep Singh Lubana***, Yongyi Yang*, Maya Okawa, Kento Nishi, Martin Wattenberg, and Hidenori Tanaka. In-context learning of representations. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2025

4. Core Francisco Park*, **Ekdeep Singh Lubana***, Itamar Pres, and Hidenori Tanaka. Competition dynamics shape algorithmic phases of in-context learning. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2025. (Accepted for **Spotlight** presentation.)

5. **Ekdeep Singh Lubana***, Kyogo Kawaguchi*, Robert P. Dick, and Hidenori. Tanaka. A percolation model of emergence: Analyzing transformers trained on a formal language. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2025

6. Abhinav Menon*, Manish Shrivastava, David Krueger, and **Ekdeep Singh Lubana***. Analyzing (in) abilities of saes via formal languages. In *Proc. North American Association of Computational Linguistics (NAACL)*, 2025. (Also **awarded best paper** at NeurIPS workshop on Foundation model interventions, 2024.)

7. Yongyi Yang, Core Francisco Park, **Ekdeep Singh Lubana**, Maya Okawa, Wei Hu, , and Hidenori. Tanaka. Dynamics of concept learning and compositional generalization. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2025

8. Core Francisco Park*, Maya Okawa*, Andrew Lee, Hidenori Tanaka, and **Ekdeep Singh Lubana**. Emergence of Hidden Capabilities: Exploring Learning Dynamics in Concept Space. In *Proc. Adv. in Neural Information Processing Systems (NeurIPS)*, 2024. (Accepted for **Spotlight** presentation.)

9. Kento Nishi, Maya Okawa, Rahul Ramesh, Mikail Khona, Hidenori Tanaka*, and **Ekdeep Singh Lubana***. Representation shattering in transformers: A synthetic study with knowledge editing. *arXiv preprint arXiv:2410.11767*, 2024

10. Pulki Gopalani, **Ekdeep Singh Lubana**, and Wei. Hu. Abrupt Learning in Transformers: A Case Study on Matrix Completion. In *Proc. Adv. in Neural Information Processing Systems (NeurIPS)*, 2024.

11. Samyak Jain, **Ekdeep Singh Lubana**, Kemal Oksuz, Tom Joy, Philip H.S. Torr, Amartya Sanyal, and Puneet K. Dokania. What Makes and Breaks Safety Fine-tuning? A Mechanistic Study. In *Proc. Adv. in Neural Information Processing Systems (NeurIPS)*, 2024.

12. Usman Anwar, Abulhair Saparov*, Javier Rando*, Daniel Paleka*, Miles Turpin*, Peter Hase*, **Ekdeep Singh Lubana***, Erik Jenner*, Stephen Casper*, Oliver Sourbut*, Benjamin Edelman*, Zhaowei Zhang*, Mario Gunther*, Anton Korinek*, Jose Hernandez-Orallo*, Lewis Hammond, Eric Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Zhang, Ruiqi Zhong, Seán Ó hÉigeartaigh, Gabriel Rachet, Giulio Corsi, Alan Chan, Markus Anderljung, Lillian Edwards, Yoshua Bengio, Danqi Chen, Samuel Albanie, Tegan Maharaj, Jakob Foerster, Florian Tramer, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*, 2024

13. Samyak Jain*, Robert Kirk*, **Ekdeep Singh Lubana***, Robert P. Dick, Hidenori Tanaka, Edward Grefenstette, Tim Rocktaschel, and David Krueger. Mechanistically Analyzing the Effects of Fine-Tuning on Procedurally Defined Tasks. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2024.

14. Eric Bigelow, **Ekdeep Singh Lubana**, Robert P. Dick, Hidenori Tanaka, and Tomer Ullman. In-Context Learning Dynamics with Random Binary Sequences. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2024.

15. Rahul Ramesh, **Ekdeep Singh Lubana**, Mikail Khona, Robert P. Dick, and Hidenori Tanaka. Compositional Capabilities of Autoregressive Transformers: A Study on Synthetic, Interpretable Tasks. In *Proc. Int. Conf. on Machine Learning (ICML)*, 2024.

16. Mikail Khona, Maya Okawa, Jan Hula, Rahul Ramesh, Kento Nishi, Robert P. Dick, **Ekdeep Singh Lubana***, and Hidenori Tanaka*. Towards an Understanding of Stepwise Inference in Transformers: A Synthetic Graph Navigation Model. In *Proc. Int. Conf. on Machine Learning (ICML)*, 2024.

17. Maya Okawa*, **Ekdeep Singh Lubana***, Robert P. Dick, and Hidenori Tanaka*. Compositional Abilities Emerge Multiplicatively: Exploring Diffusion Models on a Synthetic Task. In *Proc. Adv. in Neural Information Processing Systems (NeurIPS)*, 2023.

18. **Ekdeep Singh Lubana**, Eric J Bigelow, Robert P. Dick, David Krueger, and Hidenori Tanaka. Mechanistic Mode Connectivity. In *Proc. Int. Conf. on Machine Learning (ICML)*, 2023.

19. **Ekdeep Singh Lubana**, Johann Brehmer, Pim de Haan, and Taco Cohen. FoMo Rewards: Can we cast foundation models as reward functions? In *NeurIPS Foundation Models for Decision Making Workshop*, 2023

20. Liu Ziyin, **Ekdeep Singh Lubana**, Masahito Ueda, and Hidenori Tanaka. What Shapes the Loss Landscape of Self-Supervised Learning? In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2023.

21. Puja Trivedi and **Ekdeep Singh Lubana**, Mark Heimann, Danai Koutra, and Jay Jayaraman Thiagarajan. Analyzing Data-Centric Properties for Contrastive Learning on Graphs . In *Proc. Adv. in Neural Information Processing Systems (NeurIPS)*, 2022.

22. **Ekdeep Singh Lubana**, Ian Tang, Fahim Kawsar, Robert P. Dick, and Akhil Mathur. Orchestra: Unsupervised Federated Learning via Globally Consistent Clustering. In *Proc. Int. Conf. on Machine Learning (ICML)*, 2022. (Accepted for **Spotlight** presentation.)

23. **Ekdeep Singh Lubana**, Robert P. Dick, and Hidenori Tanaka. Beyond BatchNorm: Towards a Unified Understanding of Normalization in Deep Learning. In *Proc. Adv. in Neural Information Processing Systems (NeurIPS)*, 2021.

24. **Ekdeep Singh Lubana** and Robert P. Dick. A Gradient Flow Framework for Analyzing Network Pruning. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2021. (Accepted for **Spotlight** presentation.)

25. **Ekdeep Singh Lubana**, Puja Trivedi, Danai Koutra, and Robert P. Dick. How do Quadratic Regularizers Prevent Catastrophic Forgetting: The Role of Interpolation. In *Proc. Conf. on Lifelong Learning Agents (CoLLAs)*, 2022.

26. **Ekdeep Singh Lubana**, Robert P. Dick, Vinayak Aggarwal, and Pyari Mohan Pradhan. Minimalistic Image Signal Processing for Deep Learning Accelerators. In *Proc. Int. Conf. on Image Processing (ICIP)*, 2019.

27. **Ekdeep Singh Lubana**, Vinayak Aggarwal, and Robert P. Dick. Machine Foveation: An Application-Aware Compressive Sensing Framework. In *Proc. Data compression Conference (DCC)*, 2019.

28. **Ekdeep Singh Lubana** and Robert P. Dick. Digital Foveation: An Energy-Aware Machine Vision Framework. *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, pages 2371–2380, 2018.

Technical Service

· Reviewer for NeurIPS, ICML, ICLR, AISTATS, IEEE TPAMI, IEEE TNNLS *2021–present*
· **Top Reviewer**, NeurIPS. *2023*
· **Top Reviewer**, ICLR. *2022*
· **Top Reviewer**, NeurIPS. *2022*

Technical Awards

· Awarded the **BIRAC-GYTI award** by the **President of India.** *2018*
· Winner of the **Ericsson Innovation Challenge** held at the Nobel Museum, Stockholm, Sweden. *2017*
· Winner of the **Jury's choice award** at the **Accenture Innovation Challenge.** *2017*
· **Gold medal** and **winner of Engineers' Conclave** at **Inter-IIT Tech meet.** *2018*

Academic achievements & Scholarships

· Awarded the **KVPY (Kishore Vaigyanik Protsahan Yojna)** Fellowship by Govt. of India. *2015*
· Awarded the **NTSE (National Talent Search)** Scholarship by N.C.E.R.T., New Delhi. *2014*
· Ranked amongst **Top 300** students in **National Standard Examination in Astronomy.** *2015*
· Ranked amongst **Top 300** Students in the **Indian National Mathematics Olympiad.** *2015*