

# SAMYAK JAIN

(+91)9179144039 ◊ samyakjain.cse18@itbhu.ac.in ◊ DOB: 1<sup>st</sup> December, 1999

[LinkedIn](#) ◊ [Github](#) ◊ [Webpage](#) ◊ [Google Scholar](#) ◊ [Twitter](#)

## EDUCATION

**Indian Institute of Technology (BHU) Varanasi**

August 2018 - May 2023

Integrated Dual Degree (B.Tech + M.Tech) in Computer Science - CGPA : 9.55/10.0

[Master's Thesis](#)

## AREAS OF INTEREST

**Research topics:** AI safety, AI alignment, Science of deep learning, Interpretability, Learning dynamics, Optimization

**Sub-topics:** Adversarial robustness, Red teaming, Safety fine-tuning, Compositional generalization, Phase transitions, Mode connectivity, Domain generalization, Reward hacking, Cooperative alignment, Lottery ticket hypothesis.

## EXPERIENCE

**Microsoft Research India**

July 2024 - Present

*Research Fellow*

*Mentor* [Navin Goyal](#)

*Project:* Developing a better understanding on why lottery tickets exist using tools from interpretability.

**Five AI and Torr Vision Group, University of Oxford**

October 2023 - June-2024

*Research Intern*

*Mentor* [Puneet Dokania](#)

*Project:* Demonstrated the mechanisms involved behind the success of jailbreaking attacks.

**Krueger AI Safety Lab, University of Cambridge**

May 2023 - October-2023

*Research Intern*

*Mentor* [David Krueger](#)

*Project:* Showed that fine-tuning learns minimal transformations of a pretrained model's capabilities, like a "wrapper".

**Vision and AI Lab, Indian Institute of Science, Bangalore**

May 2020 - May-2023

*Research Intern*

*Mentor* [Venkatesh Babu](#)

*Project:* Built more effective and efficient adversarial training methods, achieving SOTA performance on leaderboards.

**Theoretical Foundations of AI Lab, Technical University of Munich**

May 2021 - August-2021

*Research Intern*

*Mentor* [Debarghya Ghoshdastidar](#)

*Project:* Worked on understanding the learning dynamics of linear autoencoders.

## PUBLICATIONS

- **What Makes Safety Fine-tuning Methods Safe? A Mechanistic Study**  
Samyak Jain, Ekdeep Singh, Kemal Oksuz, Tom Joy, Phil Torr, Amartya Sanyal, Puneet Dokania  
ICML workshop on Mechanistic Interpretability, 2024 (**Spotlight**)  
NeurIPS 2024 [main code](#)
- **Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks**  
Samyak Jain\*, Robert Kirk\*, Ekdeep Singh\*, Hidenori Tanaka, Robert Dick, Tim Rocktaschel, Edward Grefenstette, David Krueger  
ICLR 2024 [main code](#)
- **Towards Understanding and Improving Adversarial Robustness of Vision Transformers**  
Samyak Jain, Tanima Dutta  
CVPR 2024 [main](#)
- **DART: Diversify-Aggregate-Repeat Training Improves Generalization of Neural Networks**  
Samyak Jain\*, Sravanti Addepalli\*, Pawan Sahu, Priyam Dey, RV. Babu  
CVPR-2023 [main code](#)
- **Efficient and Effective Augmentation Strategy for Adversarial Training**  
Sravanti Addepalli\*, [Samyak Jain\\*](#), RV. Babu  
NeurIPS 2022 [main code](#)
- **Scaling Adversarial Training to Large Perturbation Bounds**  
Sravanti Addepalli\*, [Samyak Jain\\*](#), Gaurang Sriramanan, RV. Babu  
ECCV 2022 [main code](#)
- **Boosting Adversarial Robustness using Feature Level Stochastic Smoothing**  
Sravanti Addepalli\*, [Samyak Jain\\*](#), Gaurang Sriramanan\*, RV. Babu  
SAIAD Workshop CVPR 2021 [main code](#)

## FEATURED ACADEMIC PROJECTS AND COLLABORATIONS

---

### Understanding the lottery ticket hypothesis [Navin Goyal](#)

- Observed that neurons forming lottery tickets have a high projection with the final model at initialization itself.
- High projection leads to exponential rise in the norm of such neurons, which leads to faster convergence.

### Mechanistic understanding of safety fine-tuning and jailbreaking attacks [Puneet Dokania](#), [Ekdeep Singh](#), [Amartya Sanyal](#), [Phil Torr](#)

- Observed that safety fine-tuning projects unsafe samples into model's null space, thereby leading to safe behavior.
- Demonstrated that the learned projection is low ranked in nature, which makes it easy to craft jailbreaks.
- [Gemma Scope](#) highlighted the using sparse autoencoders based on insights in this work could help improve safety.

### Mechanistic understanding of fine-tuning [Robert Kirk](#), [Ekdeep Singh](#), [David Krueger](#), [Hidenori Tanaka](#), [Tim Rocktaschel](#), [Edward Grefenstette](#)

- Demonstrated that fine-tuning is unable to alter the model mechanistically, but rather gives a pretense of change.
- Proposed reverse fine-tuning to demonstrate this, and it has now become a staple method to evaluate unlearning.
- [Follow-up](#) works have used key insights from our work to counter use of safety fine-tuning as an assurance protocol.

### Exploring loss basin to find generalized solutions [RV. Babu](#), [Sravanti Addepalli](#)

- Derived bounds showing that weight averaging of diverse models in training slows learning of spurious features.
- Proposed method DART demonstrated improvements on both in-domain and out of domain settings.

### Using data augmentations effectively in adversarial training [RV. Babu](#), [Sravanti Addepalli](#)

- Showed for the first time that it is possible to use augmentations effectively and efficiently in adversarial training.
- Demonstrated that weight space smoothing with single-step attacks can help in preventing catastrophic overfitting.

### Aligning adversarial training with Ideal training objectives [RV. Babu](#), [Sravanti Addepalli](#)

- Observed that standard AT methods fail to generalize to larger perturbation bounds due to conflict in training.
- Proposed a method, which aims to align the model's predictions with the oracle labels of adversarial images.

### Understanding gradient masking in vision transformers [Tanima Dutta](#)

- Past works have demonstrated gradient masking in vision transformers, but failed to understand the cause.
- Demonstrated that softmax in attention causes floating point errors, which leads to gradient masking in ViTs.

## SCHOLASTIC ACHIEVEMENTS

---

- Recipient of **DAAD-WISE**, a research oriented scholarship program by German Government.
- Fellow of Berkeley Existential Risk Initiative (**BERI**), which supported my research at Cambridge.
- Recipient of Summer Research Fellowship 2020 (**SRFP**), a research program by Indian Government.
- All India rank 922 in JEE Advanced 2018 and 346 in JEE Mains 2018 out of 1 million+ candidates.
- Selected for the KVPY 2018 Fellowship (IISc, Bangalore) by the Govt. of India.
- Ranked in amongst **Top 300** students in India for Maths, Physics and Astronomy Olympiads at national level – INMO, INPhO, INAO 2018. City topper in National Talent Search Exam (NTSE) 2016.
- Member of [Future of Life-Existential AI Safety Community](#).

## INVITED TALKS AND PRESENTATIONS

---

<b>Mechanistic understanding of safety fine-tuning and jailbreaks</b> ICML mechanistic interpretability workshop	July 2024
<b>Pitfalls in safety fine-tuning for robust alignment</b> ETH Zurich AI Center.	February 2024
<b>Mechanistic understanding of fine-tuning</b> Krueger AI safety lab, University of Cambridge and Five AI.	November 2023

## FEATURED POSITIONS AND RELEVANT COURSES

---

**Reviewer:** NeurIPS 2024, ICLR 2024, ICML 2023, NeurIPS 2023, CVPR 2023, CVPR 2022, ICLR 2022, NeurIPS 2022.

**Outstanding / Highlighted reviewer award:** NeurIPS 2024, CVPR 2023, CVPR 2022, ICLR 2022

**Teaching Assistant:** Introduction to Database Management and Introduction to Machine Learning

- Conducted lab classes of undergraduate students with a batch size of over eighty students.
- Worked alongside the professor to design and evaluate lab assignments and final course assessments.

**Relevant Courses:** Computer Vision (**A**), Applications of Deep Learning (**A**), Theory of computation (**A-**), Artificial Intelligence (**A**), Probability and Statistics (**A**), Stochastic processes (**A**), Linear Algebra (**A**), Data Mining (**A**), Computer Graphics (**A\***), Calculus (**A**), Signal Processing (**A**), Number Theory (**A-**), Data Structures (**A-**) and Algorithms (**A\***), Information Security (**A\***), Rings and Modules (**A**), Probabilistic Graphical Models and Optimization (online).