Nishad **Singhi**

University of Tübingen & Max Planck Institute for Intelligent Systems

nishadsinghi.github.io @ nishadsinghi@gmail.com github.com/nishadsinghi Google Scholar

Education

Present | **University of Tübingen** — **GPA: "Very Good" | 3.76/4 (US Scale)**
2020 | MSc in Neural Information Processing

2020 | **Indian Institute of Technology (IIT) Delhi** — **GPA: 8.6/10**
2016 | BTech in Electrical Engineering *(Specialization in Cognitive and Intelligent Systems)*

Research Interests

Multimodal Learning, Robust and Explainable AI, Representation Learning, Computational Cognitive Science

Publications

C = Conference, R = Report

[C.1] **CleanCLIP: Mitigating Data Poisoning Attacks in Multimodal Contrastive Learning** [Paper]
Hritik Bansal*, Nishad Singhi*, Yu Yang, Fan Yin, Aditya Grover, Kai-Wei Chang (* = Equal Contribution)
International Conference on Computer Vision (ICCV) 2023 **(Oral; Top 1.8%)**
Best Paper Award at the RTML Workshop at ICLR 2023 [ICCV 2023]

[C.2] **Toward a normative theory of (self-)management by goal-setting** [Paper]
Nishad Singhi, Florian Mohnert, Ben Prystawski, Falk Lieder
Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci) 2023 **(Oral)**
Diversity and Inclusion Award (10 recipients worldwide) [CogSci 2023]

[C.3] **Using Computational Models to Understand the Role and Nature of Valuation Bias in Mixed Gambles** [Paper]
Nishad Singhi, Sumeet Agarwal, Sumitava Mukherjee
Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci) 2023 [CogSci 2023]

[C.4] **An fMRI Study of Goal-Directed Behaviour under Approach and Avoidance Goals** [Paper] [Poster]
Nishad Singhi, Michiko Sakaki, Kou Murayama, et al.
Psychologie und Gehirn (PuG) 2023 [PuG 2023]

[R.1] **Computational Principles of Metacognitive Reinforcement Learning** [Paper]
Nishad Singhi, *Survey 2022*

Select Research Projects

CleanCLIP: Defending CLIP Against Backdoor Attacks [🌐] Nov'22 - Apr'23
Advisors: Prof. Kai-Wei Chang, Prof. Aditya Grover (UCLA Computer Science)

- Objective: Defend Multimodal Contrastive Models (e.g., CLIP) against backdoor attacks.
- Proposed a lightweight, unsupervised technique to neutralize backdoors from a poisoned CLIP model. Our method finetunes a poisoned model on a small amount of clean data to learn the modalities of images and texts independently. This breaks the spurious correlation between the backdoor trigger and the target label, purifying the model.
- Demonstrated the effectiveness of our technique against a wide range of attacks while maintaining task performance.

Enhancing Mechanistic Interpretability in Neural Networks Nov'22 - Present
Advisors: Dr. Wieland Brendel (MPI for Intelligent Systems)

- Objective: Build Neural Networks wherein every neuron activates for a specific concept, thereby enhancing their mechanistic interpretability.
- We associate each neuron with a specific concept represented by a point in the CLIP embedding space. Then, we train the network to position highly activating images close to the concept descriptor within the CLIP embedding space.

Intervention Friendly Concept-Bottleneck Models Apr'23 - Present
Advisor: Prof. Zeynep Akata (University of Tübingen)

- Objective: Allow humans to correct the predictions of Concept-Based Models in a sample-efficient manner.
- We model relationships between individual concepts using Graph Neural Nets, so that the model can automatically update the predictions of all concepts when an expert corrects any concept.
- Exploring the use of Reinforcement Learning to decide which concept should be verified by the expert.

Automatic Subgoal Discovery for Goal Achievement [🌐] Mar'21 - Mar'23

Advisor: Dr. Falk Lieder (MPI for Intelligent Systems)

- Objective: Automatically decompose a challenging problem into easier subgoals to improve people's performance.
- Developed a theoretical framework to derive the subgoals that best improve people's performance on a task.
- Employed a cognitive model to simulate people's actions given a goal and subgoal. Then, used optimization techniques to compute subgoals with the largest performance improvement.
- Demonstrated via behavioral experiments that people with our subgoals perform better and use 3x fewer resources.

fMRI Study of Motivation under Approach and Avoidance Goals [🌐] Dec'21 - Feb'22

Advisors: Prof. Kou Murayama, Prof. Michiko Sakaki (University of Tübingen)

- Objective: Understand how the brain processes *Approach ("achieve success")* and *Avoidance ("avoid failure")* goals.
- People enjoyed approach tasks and felt anxious in avoidance tasks. We found no differences in the brain's reward circuit.

Computational Modeling of Loss Aversion [🌐] Jul'19 - Jul'20

Advisors: Prof. Sumeet Agarwal, Prof. Sumitava Mukherjee (IIT Delhi)

- Objective: Understand why humans dislike gambles that can result in a loss (e.g., win $11 or lose $10 with equal prob.).
- Employed computational models of decision-making to show that a valuation bias affects people's choices and a prior bias to reject affects response times. Demonstrated that valuation bias may be linked to attentional mechanisms.

Modeling Social Perception in Physical Domains May'19 - July'19

Advisor: Prof. Tao Gao (UCLA Statistics)

- Objective: Model how humans infer the intention of physical agents by observing their actions.
- Built a generative model of agents' actions conditioned on their intent in MuJoCo using Deep Reinforcement Learning.

Honours and Awards

Best Paper Award ($1,000), 2023 [🌐] as co-first author for CleanCLIP at the RTML workshop, *ICLR 2023*.

Diversity and Inclusion Award ($1,000), 2023 [🌐] Among 10 recipients worldwide awarded at *CogSci 2023*.

Bounded Rationality Winter School, 2020 Among 40 selected worldwide for winter school organized by MPI Berlin.

Prof. R. K. Mittal Award (INR 10,000), 2017 Awarded to 2 freshmen (out of 850+) at IIT Delhi for academic performance.

IIT Delhi Merit Award (INR 2,500), 2017 Conferred for being among the top 7% students of the batch at IIT Delhi.

IIT-JEE, 2016 Ranked amongst the top 0.01% applicants out of 1.5 million candidates in the IIT-JEE entrance exam.

Talks

Toward a normative theory of (self-)management by goal-setting

- The 44th Annual Meeting of the Cognitive Science Society July 2023

CleanCLIP: Mitigating Data Poisoning Attacks in Multimodal Contrastive Learning

- Trustworthy and Reliable Large-Scale Machine Learning Models Workshop at ICLR 2023 May 2023

Teaching

Teaching Assistant, Introduction to Computational Neuroscience *University of Tübingen* Winter 2022

Teaching Assistant, Introduction to Biology for Engineers *IIT Delhi* Spring 2020

Relevant Coursework

- **Machine Learning**: Computer Vision, NLP, Explainable ML, Probabilistic Machine Learning, Deep Learning
- **EE & CS**: Data Structures and Algorithms, Information Theory, Signal Processing
- **Neuroscience**: Neural Dynamics, Neural Coding, Neural Data Analysis, Computational Motor Control

Leadership and Volunteering

Student Affairs Council IIT Delhi, 2019 As a member of the apex student body at IIT Delhi, I was involved in policy-making and taking initiatives to solve student-related problems.

Teaching Volunteer, Ibtada, 2017 Spent a summer teaching English and basic computer skills to underprivileged girls.