# Nishad Singhi

nishad.singhi@student.uni-tuebingen.de |


## Education

**University of Tübingen** 2020 - Present
Master of Science in Neural Information Processing Current GPA: *Excellent* | 3.76/4 (US Scale) | 1.24/1 (German Scale)

**Indian Institute of Technology Delhi** 2016 - 2020
Bachelor of Technology in Electrical Engineering GPA: 8.6/10
*Specialization in Cognitive and Intelligent Systems*

## Publications & Reports

**CleanCLIP: Mitigating Data Poisoning Attacks in Multimodal Contrastive Learning**
Hritik Bansal*, **Nishad Singhi***, Yu Yang, Fan Yin, Aditya Grover, Kai-Wei Chang. *ICCV 2023.* ***(Oral;Top 2%)*** [Paper]
***Best Paper Award*** *at the RTML Workshop, International Conference for Learning Representations (ICLR), 2023.*

**Toward a normative theory of (self-)management by goal-setting**
**Nishad Singhi**, Florian Mohnert, Ben Prystawski, Falk Lieder. *CogSci 2023.* ***(Oral)*** [Paper]
***Diversity & Inclusion Award*** *at the Annual Meeting of the Cognitive Science Society (1/10 recipients worldwide).*

**Using Computational Models to Understand the Role and Nature of Valuation Bias in Mixed Gambles**
**Nishad Singhi**, Sumeet Agarwal, Sumitava Mukherjee. *CogSci 2023.* [Paper]

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

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

## Selected Experience

**CleanCLIP: Defending CLIP Against Backdoor Attacks** Prof. Kai-Wei Chang & Prof. Aditya Grover, UCLA
*UCLA NLP* *Nov 2022 - Mar 2023*

- Objective: Defending multimodal contrastive learning models (e.g., CLIP) against backdoor attacks.
- Demonstrated the effectiveness of combining multi-modal contrastive loss (e.g., CLIP) and uni-modal self-supervision loss (e.g., SimCLR) for fine-tuning on limited clean data, countering attacks while maintaining high downstream performance.

**Enhancing Mechanistic Interpretability in Neural Networks** Dr. Wieland Brendel, MPI-IS
*Robust Machine Learning Group, MPI for Intelligent Systems* *Nov 2022 - Present*

- Objective: Develop a regularization approach for training neural networks that ensures each neuron aligns with a distinct semantic concept, thereby enhancing the mechanistic interpretability of deep neural networks.
- We associate each neuron with a specific concept represented by a point in the CLIP embedding space. Then, we train the neural network to position highly activating images close to the concept descriptor within the CLIP embedding space.

**Automatic Subgoal Discovery for Goal Achievement** Dr. Falk Lieder, MPI-IS
*Rationality Enhancement Group, MPI for Intelligent Systems* *Mar 2021 - Present*

- Objective: Create a theoretical framework for setting subgoals that can assist individuals in achieving their objectives.
- We use a computational model of human behavior to estimate how much a subgoal benefits individuals. Then, we use optimization methods to identify subgoals that optimize performance. Our subgoals improve performance in user studies.

**Multi-Agent Reinforcement Learning in Physical Environments** Prof. Tao Gao, UCLA
*Visual Intelligence Lab, UCLA* *May 2019 - Jul 2019*

- Objective: Learn policies for multi-agent adversarial games in MuJoCo using Deep Reinforcement Learning.
- Implemented AlphaZero and benchmarked its performance against model-free, tree-search algorithms, and heuristic policies.

## Achievements and Honours

- **Best Paper Award** ($1,000) as a co-first author for CleanCLIP at the RTML Workshop at ICLR, 2023. [2023]
- **Diversity & Inclusion Award** ($1,000): Among 10 recipients (worldwide) awarded by Cognitive Science Society. [2023]
- **Bounded Rationality Winter School:** Among 40 selected (worldwide) for winter school organized by MPI Berlin. [2020]
- **Prof. RK Mittal Award** (INR 10,000): Given to 2 freshmen (out of 850) at IITD for academic performance. [2017]
- **IIT Delhi Merit Award:** Conferred for securing a position among the top 7% students of the batch at IIT Delhi. [2017]
- **IIT-JEE:** Ranked amongst the top 0.01% applicants out of 1.5 million candidates in IIT-JEE Advanced. [2016]