




Shashwat Goel

 <https://shash42.github.io>
 shashwatnow@gmail.com
 <https://github.com/shash42>

Education

- 2024- **PhD in Artificial Intelligence**
ELLIS Institute and Max Planck Institute for Intelligent Systems, Tübingen
ADVISED BY: Jonas Geiping and Douwe Kiela
- 2019-2024 **B.Tech. and M.S. (by Research) in Computer Science Engineering**
International Institute of Information Technology (IIIT), Hyderabad GPA: 9.60/10
THESIS: **New Frontiers for Machine Unlearning**, advised by Prof. Ponnurangam K.

Research Experience

- July-Dec 2023 **Researcher**, *Stanford Existential Risk Institute ML Alignment Theory Scholars (SERI MATS)*
MENTOR: Dan Hendrycks
- May-June 2023 **Quantitative Research Intern**, *Central Research Team, Millennium India*
PROJECT: AutoML for Tree-based and linear ensembles to find alpha across datasets.
- May-July 2022 **Research Intern**, *Social Choice Theory, LAMSADE, CNRS*
ADVISORS: Jerome Lang, Dominik Peters
- July-Sept 2021 **Research Assistant**, *Language Evolution, Santa Fe Institute*
MENTOR: Tanmoy Chakroborty
- May-June 2021 **Developer**, *Distributed Computing Laboratory, Summer@EPFL*
MENTOR: Matteo Monti, Rachid Guerraroui
- April-Aug 2020 **Research Developer**, *Apertium, Google Summer of Code*
MENTORS: Mikel Forcada, Jorge Gracia

Publications

- [10] **Corrective Machine Unlearning**
Shashwat Goel*, Ameya Prabhu*, Philip Torr, P. Kumaraguru, Amartya Sanyal
Recommended for DMLR Journal (top 15/100+ submissions) at the *Workshop on Data-centric Machine Learning Workshop (DMLR)*
12th International Conference on Representation Learning (ICLR) 2024 [paper]
- [9] **The WMDP Benchmark: Measuring and Reducing Malicious Use with Unlearning**
Center for AI Safety
International Conference on Machine Learning (ICML) 2024 [paper]
- [8] **Proportional Aggregation of Preferences for Sequential Decision Making**
Nikhil Chandak, Shashwat Goel, Dominik Peters
Outstanding Paper Award (top 3 out of 12,000+ submissions) at the *38th Annual Conference of the Association for the Advancement of Artificial Intelligence (AAAI) 2024* [paper]
- [7]

Representation Engineering: A Top-Down Approach to AI Transparency

Center for AI Safety

ArXiv 2023

[\[paper\]](#)

- [6] ***Probing Negation in Language Models***
Shashwat Singh*, **Shashwat Goel***, Saujas Vaduguru, Ponnurangam Kumaraguru
8th Workshop on Representation Learning for NLP (RepLANLP)
61st Annual Meeting of the Association for Computational Linguistics (ACL) 2023 [\[paper\]](#)
- [5] ***Towards Adversarial Evaluations of Inexact Machine Unlearning***
Shashwat Goel*, Ameya Prabhu*, Amartya Sanyal, Ser-Nam Lim, Phillip Torr, Ponnurangam Kumaraguru
ArXiv 2023 [\[paper\]](#)
- [4] ***Low Impact Agency: Review and Discussion***
Danilo Naiff, **Shashwat Goel**
ArXiv 2022 [\[paper\]](#)
- [3] ***Modelling and Optimizing the Allocation of COVID-19 Swabs to Labs***
Nikhil Chandak, **Shashwat Goel**, Kunal Jain, Arpan Dasgupta
Student Abstract at 18th Mixed Integer Programming Workshop 2021
Winner, Covid-19 Swabs2Labs Hackathon by Ministry of Health Karnataka [\[paper\]](#)
- [2] ***Bilingual Dictionary Generation and Enrichment via Graph Exploration***
Shashwat Goel, Jorge Gracia, Mikel L. Forcada
Special Issue on *Latest Advancements in Linguistic Linked Data* 2021
Semantic Web Journal [\[paper\]](#)
- [1] ***From Pivots to Graphs: Augmented Cycle Density as a generalization to One Time Inverse Consultation***
Shashwat Goel, Kunwar Shanjeet Grover
4th Shared Task on Translation Inference Across Dictionaries
3rd Conference on Language, Data and Knowledge 2021 [\[paper\]](#)

Honours and Awards

- 2024 Outstanding Paper Award (Top 3/12,000+), AAI Conference in Vancouver, Canada
Outstanding Reviewer (Top 10%): ICML 2022, ICLR DMLR Workshop 2024
- 2020 Finalist (Top 50/3000+), ACM-ICPC Indian Regionals
- 2019 Honorable Mention, International Olympiad of Linguistics
- 2019 National Rank 6, International Olympiad of Informatics Indian Team Selection
- 2017 Grand Prize Winner (1/1500+), NASA Ames Space Settlement Design Contest

Teaching Experience

- Spring 2024 **Head Teaching Assistant**, [Responsible and Safe AI](#), IIIT Hyderabad
- Spring 2023 **Facilitator**, [AI Safety Fundamentals](#), BlueDot Impact
- Spring 2023 **Teaching Assistant**, Topics in DL (Graph Neural Networks), IIIT Hyderabad
- Fall 2022 **Teaching Assistant**, [Automata Theory](#), IIIT Hyderabad

Academic Service and Outreach

Reviewer: CoLLAs 2024, ICLR DMLR Workshop 2024, AISTATS 2024, CoLLAs 2023, CODS-COMAD 2023, ICML 2022

University Groups: ML Reading Group (Founder), Effective Altruism Group (Founder), Theory Group (Former Head), Programming Club (Former Head), Parliamentary Debate Team, Student Magazine (Editor)

Trainer: Indian Team Selection for the International Olympiad of Informatics 2020, Panini Linguistics Olympiad 2024

Talks: How can we deal with Conflicting Training Signal in Deep Learning, Intro to DL Research, Pathways from Cognition to DL Research, Voting Rules and Fair Representation, Perfect Information Sequential Games, Graph Theory for high-schoolers, Intro to Effective Altruism, Making Calibrated Predictions, Population Ethics, Linguistics Olympiad Training Workshops
