Syed Ali Haider

syed.ali.haider.gr@dartmouth.edu · s-haider10.github.io · github.com/s-haider10 · linkedin/haider-tech

Education

Dartmouth College Sep 2025 – Jun 2027 *(expected)*
M.S. in Computer Science (Research Track) *GPA: 4.0 / 4.0*
Selected coursework: Mechanistic Interpretability, Neurosymbolic AI, AI Agents, Multimodal AI, Human–AI Interaction

New York University Aug 2021 – May 2025
B.S. in Computer Science, Minor in Mathematics

Recent Experience

Minds, Machines & Society Group, Dartmouth College Jan 2026 – Present
Graduate Researcher | Advisor: Prof. Soroush Vosoughi

- Conducted systematic study of how misaligned minorities reshape collective belief formation in multi-agent LLM debate (27,900 trials, 4 network topologies); evidence that debate-based scalable oversight is non-robust to adversaries.
- Designed value-installation framework testing whether prompt-installed moral identities (deontologist / utilitarian / virtue ethicist) hold under sustained adversarial pressure in iterated Prisoner's Dilemma; pilot shows agents rationalize defection within the framework rather than abandon it.

Science & Art of Human-AI Systems Lab, Dartmouth College Jan 2026 – Present
Graduate Researcher | Advisor: Prof. Nikhil Singh

- Tested communication and deception in repeated PD with private signals across 4 alignment tiers (Base / SFT / RL / HHH) on Qwen-2.5-32B; 864 agent-rounds show no detectable effect of tier ($\chi^2 = 1.07$, $p = 0.78$), with prosocial framing dominating regardless of alignment.
- Identified topology-dependent correction failure in multi-agent networks: scale-free communication structures fail to self-correct after misinformation injection, while other topologies recover.

Publications & Preprints

[1] *Policy Extraction and Enforcement for Runtime AI Governance.* **Accepted** — ICML Technical AI Governance '26
S. A. Haider, B. Huh, H. H. Kim, G. Nalagatla, C. Guerrero Alvarez, Y. Raj, S. Vosoughi.

[2] *Misalignment Contagion: How a Minority Shifts Aligned LLM Agents in Debate.* Under review — COLM '26
S. A. Haider, S. Vosoughi.

[3] *Position: Belief Attributions Require Behavioral Evidence.* Under review — ICML Philosophy x ML '26
S. A. Haider.

[4] *Position: What is a Morally Aligned AI Agent?* Under review — ICML Philosophy x ML '26
S. A. Haider.

[5] *Log-Probability Guided Adversarial Attacks on Multi-Agent LLM Debate.* Manuscript under preparation '26
F. Niyigaba*, **S. A. Haider***, Nikhil Singh.

[6] *Harnessing Multilingual Geometry for Knowledge Transfer.* Manuscript in preparation for ACL ARR Aug '26
F. Niyigaba*, **S. A. Haider***, J. Lee, S. Vosoughi.

Honors & Awards

Thomas D. Sayles Research Grant ($2,000), Dartmouth Ethics Institute 2026
Recognition of Excellence Citation, Dartmouth College *(awarded twice)* 2026
Best Poster, Dartmouth Technigala 2025
Gaurini Merit Scholarship ($50,000), Dartmouth College 2025
Gobi Partners VC Fellow — 1 of 12 selected from 2,000 applicants 2025
Dean's Undergraduate Research Fund ($3,500), NYU 2023 & 2024
Best Research Poster, NYU CS Undergraduate Research Symposium 2024

Teaching & Service

Reviewer ICML Workshops '26: Agents in the Wild, Trustworthy AI for Good, Philosophy Meets ML, Technical AI Governance.
Dartmouth Graduate TA (2025–26): Machine Learning, Deep Learning, Artificial Intelligence.
NYU Lead CS TA (2022–25): Machine Learning, Data Structures, Intro to Computer & Data Science.

Technical Skills

ML / AI stack: Python, PyTorch, HuggingFace, TransformerLens, Weights & Biases, vLLM, Tinker, NetworkX.
Methods: LoRA, GRPO, PPO, DPO, agentic systems (ReAct, multi-agent debate), red-teaming & UKAISI Inspect evaluation, mechanistic interpretability (logit lens, activation patching, linear probing, SAE), game theory.