Baturay Saglam

🌐 baturaysaglam.com   baturaysaglam

Education

Yale University, New Haven, CT *Aug 2023 – Present*
Ph.D. in Electrical and Computer Engineering
Advisor: Dionysis Kalogerias

Bilkent University, Ankara, Turkey *2020 – 2023*
M.S. in Electrical and Electronics Engineering
Thesis: Novel Deep RL Algorithms for Continuous Control

Bilkent University, Ankara, Turkey *2016 – 2020*
B.S. in Electrical and Electronics Engineering

Experience

Cisco Foundation AI, San Francisco, CA *Mar 2025 – Aug 2025*
Research Intern
Foundation-model training: pretraining, instruction tuning, and reasoning.

Robust Intelligence *(acquired by Cisco)*, San Francisco, CA *Sep 2024 – Mar 2025*
Research Intern
LLM robustness and AI safety: prompt-injection firewall development and evaluation.

Google Research, New York, NY *Jun 2024 – Sep 2024*
Research Intern
OOD evaluation and length generalization in autoregressive LMs.

Turk Telekom, Ankara, Turkey *May 2022 – Jul 2023*
Research Ops Specialist
RL for 5G/6G wireless control and resource allocation.

Publications

Preprints and Working Papers

[W2] **Baturay Saglam** and Dionysis Kalogerias. “Self-Improving In-Context Learning: A Test-Time Calibration with Only a Few More Forward Passes.” 2026. *(work in progress)*

[W1] **Baturay Saglam** and Dionysis Kalogerias. “Compatible Gradient Approximations for Actor-Critic Algorithms.” 2025. *(major revision in progress)* [arXiv] [code]

[P2] **Baturay Saglam** and Dionysis Kalogerias. “Test-Time Detoxification Without Training or Learning Anything.” 2026. *(under review)* [code]

[P1] Supriti Vijay, Aman Priyanshu, Anu Vellore, **Baturay Saglam**, Amin Karbasi. “Think Before You Retrieve: Learning Test-Time Adaptive Search with Small Language Models.” 2026. *(under review)* [arXiv]

Last updated: 2026-01-09

Journal Papers

[J4] **Baturay Saglam**, Dogan C. Cicek, Furkan B. Mutlu, Suleyman S. Kozat. “Mitigating Off-Policy Bias in Actor-Critic Methods with One-Step Q-learning: A Novel Correction Approach.” *Transactions on Machine Learning Research*, 2024. [paper] [code]

[J3] **Baturay Saglam**, Furkan B. Mutlu, Dogan C. Cicek, Suleyman S. Kozat. “Parameter-free Reduction of the Estimation Bias in Deep Reinforcement Learning for Deterministic Policy Gradients.” *Neural Processing Letters*, 2024. [paper]

[J2] **Baturay Saglam**, Furkan B. Mutlu, Dogan C. Cicek, Suleyman S. Kozat. “Actor-Prioritized Experience Replay.” *Journal of Artificial Intelligence Research*, 2023. [paper] [code]

[J1] **Baturay Saglam** and Suleyman S. Kozat. “Deep Intrinsically Motivated Exploration in Continuous Control.” *Machine Learning*, 2023. [paper] [code]

Conference Papers

[C7] **Baturay Saglam**, Paul Kassianik, Blaine Nelson, Sajana Weerawardhena, Yaron Singer, Amin Karbasi. “Large Language Models Encode Semantics and Alignment in Linearly Separable Representations.” *International Joint Conference on Natural Language Processing and the Asian Chapter of ACL*, 2025. [arXiv] [code]

[C6] Jane H. Lee, **Baturay Saglam**, Spyridon Pougkakiotis, Amin Karbasi, Dionysis Kalogerias. “Risk-Averse Constrained Reinforcement Learning with Optimized Certainty Equivalents.” *NeurIPS*, 2025. [paper] [code]

[C5] **Baturay Saglam**, Xinyang Hu, Zhuoran Yang, Dionysis Kalogerias, Amin Karbasi. “Learning Task Representations from In-Context Learning.” *Findings of ACL*, 2024. [paper] [code]

[C4] **Baturay Saglam**, Doga Gurgunoglu, Suleyman S. Kozat. “Deep Reinforcement Learning Based Joint Downlink Beamforming and RIS Configuration in RIS-Aided MU-MISO Systems Under Hardware Impairments and Imperfect CSI.” *IEEE International Conference on Communications Workshops*, 2023. [paper] [code]

[C3] Dogan C. Cicek, Enes Duran, **Baturay Saglam**, Kagan Kaya, Furkan B. Mutlu, Suleyman S. Kozat. “AWD3: Dynamic Reduction of the Estimation Bias.” *IEEE International Conference on Tools with Artificial Intelligence*, 2021. [paper]

[C2] Dogan C. Cicek, Enes Duran, **Baturay Saglam**, Furkan B. Mutlu, Suleyman S. Kozat. “Off-Policy Correction for Deep Deterministic Policy Gradient Algorithms via Batch Prioritized Experience Replay.” *IEEE International Conference on Tools with Artificial Intelligence*, 2021. [paper]

[C1] **Baturay Saglam**, Enes Duran, Dogan C. Cicek, Furkan B. Mutlu, Suleyman S. Kozat. “Estimation Error Correction in Deep Reinforcement Learning for Deterministic Actor-Critic Methods.” *IEEE International Conference on Tools with Artificial Intelligence*, 2021. [paper]

Model Development and Open-Source Software

Latent-Space AI Firewall 🔗 *2025*

Creator

An LLM guardrail that operates on the hidden states of Foundation-Sec-8B-Instruct; outperforms text-level filters (e.g., Llama Guard 3-8B) at detecting malicious inputs and adversarial prompts.

Foundation-Sec-8B-Instruct 🔗 *2025*

Core Model Developer

Built and released the instruction-tuned (+RLHF) variant of Foundation-Sec-8B; achieved state-of-the-art results on cybersecurity benchmarks and instruction-following tasks.

Foundation-Sec-8B 🔗 *2025*
Core Model Developer
Developed and released Cisco's cybersecurity-focused foundation model (pretrained transformer); achieved state-of-the-art performance on cybersecurity benchmarks versus comparable-size models.

FAITH: Foundation-AI Testing Hub for Cybersecurity 🔗 *2024 – 2025*
Co-Creator
Open-source evaluation harness for benchmarking language-model knowledge and capabilities in cybersecurity.

Deep RL for Joint Beamforming and Phase-Shift Optimization (RIS-MISO) ★ 208 *2021*
Creator and Maintainer
Python framework for optimizing RIS-assisted multiuser MISO systems via deep RL, targeting 6G wireless settings.

Honors and Awards

PhD Fellowship, University of California, San Diego *2023*

PhD Fellowship, Yale University *2023*

Graduate Fellowship, Turk Telekom & Information and Communication Technologies Authority *2022*

Graduate Fellowship, Directorate of Science Fellowships and Grant Programmes, Scientific and Technological Research Council of Turkey *2021*

Scholarship and Governmental Grant (up to $293K), Scientific and Technological Research Council of Turkey, Directorate of Research Support Programs *2020*

Merit-Based Scholarship, Bilkent University *2020*

Merit-Based Scholarship, Bilkent University *2016*

Teaching and Academic Service

Program Committee Member, *NeurIPS Workshop on Reliable Machine Learning* *2025*

Graduate Teaching Assistant, *CPSC 1100 — Python Programming for Humanities and Social Sciences, Yale University* *Fall 2025*

Graduate Teaching Assistant, *S&DS 617 — Advances in Large Language Models: Theory and Applications, Yale University* *Spring 2025*

Graduate Teaching Assistant, *CS 115 — Introduction to Programming in Python, Bilkent University* *Spring 2023*

Reviewer, *ICLR, NeurIPS, Machine Learning, ICASSP, IEEE Robotics and Automation Letters, ACM/IEEE HRI, IEEE RO-MAN* *2022 – Present*

Last updated: 2026-01-09