Usman Anwar

✉ usmananwar391@gmail.com

🏠 https://uzman-anwar.github.io/

in www.linkedin.com/in/uzman-anwar

Education

University of Cambridge, Cambridge, UK October 2022 - July 2026
PhD In Engineering (Machine Learning)
Awards: Open Phil AI Fellowship and Future of Life Fellowship in AI Existential Safety

Information Technology University, Lahore September 2019 – August 2021
MS Data Science (Rank: 2*nd*, Dean's Honours List)
Thesis Paper: Inverse Constrained Reinforement Learning (*published at ICML*)

University of Engineering and Technology, Lahore August 2015 – May 2019
BS Electrical Engineering
Undergraduate Thesis/FYP: Single Channel Acoustic Source Separation And Speech Enhancement

Awards & Honours

Open Phil AI Fellowship by Open Philanthrophy Foundation (2022-2026). 🔗

Vitalik Buterin PhD Fellowship by Future Of Life Institute (2022-2026). 🔗

Free Registration Award at virtual MLSS 2021 Taipei.

Graduate Student Fellowship, ITU, Lahore.

Merit Scholarship, ITU, Lahore.

Preprints

** denotes equal contribution*

U. Anwar, J. von Oswald, L. Kirsch, D. Krueger, and S. Frei. Adversarial robustness of in-context learning in transformers. 2024b. *Under review at ICLR*

U. Anwar*, A. Pandian*, J. Wan, D. Krueger, and J. N. Foerster. Noisy zero-shot coordination: Breaking the common knowledge assumption in zero-shot coordination games. 2024. *Under review at AAMAS*

T. Bush, **U. Anwar***, S. Chung*, A. Garriga-Alonso, and D. Krueger. Interpreting emergent planning in model-free reinforcement learning. 2024. *Under review at ICLR*

M. Farrugia-Roberts, K. A. Abdel Sadek*, H. Erlebach, C. Schroeder de Witt, D. Krueger, **U. Anwar**, and M. D. Dennis. Mitigating goal misgeneralization via minimax regret. 2024. *Under review at ICLR*

Conference & Journal Publications

U. Anwar, A. Saparov, J. Rando, D. Paleka, et al. Foundational challenges in assuring alignment and safety of large language models. *Transactions on Machine Learning Research*, 2024a. URL https://openreview.net/forum?id=oVTkOs8Pka. Survey Certification, Expert Certification

T. Coste, **U. Anwar**, R. Kirk, and D. Krueger. Reward model ensembles help mitigate overoptimization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=dcjtMYkpXx

D. Papadimitriou, **U. Anwar**, and D. S. Brown. Bayesian methods for constraint inference in reinforcement learning. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=oRjk5V9eDp

U. Anwar*, S. Malik*, A. Aghasi, and A. Ahmed. Inverse constrained reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. URL https://proceedings.mlr.press/v139/malik21a.html. Spotlight.

Workshop Publications

J. M. L. Rangel, **U. Anwar**, S. Schoepf, J. Foster, and D. Krueger. Learning to forget using diffusion hypernetworks. In *The Third Workshop on New Frontiers in Adversarial Machine Learning*, 2024. URL https://openreview.net/forum?id=vvRQx8oXqX

M. Brumley, J. Kwon, D. Krueger, D. Krasheninnikov, and **U. Anwar**. Comparing bottom-up and top-down steering approaches on in-context learning tasks. In *MINT: Foundation Model Interventions*, 2024. URL https://openreview.net/forum?id=VMiWNaWVJ5

A. Chan, N. Kolt, P. Wills, **U. Anwar**, C. S. de Witt, N. Rajkumar, L. Hammond, D. Krueger, L. Heim, and M. Anderljung. IDs for AI systems. In *NeurIPS 2024 Workshop on Regulatable ML*, 2024. URL https://openreview.net/forum?id=orQGtt5orZ

A. Clark, S. A. Siddiqui, R. Kirk, **U. Anwar**, S. Chung, and D. Krueger. Domain generalization for robust model based offline reinforcement learning. In *NeurIPS 2022 Workshop on Distribution Shifts and Offline Reinforcement Learning*, 2021

S. Malik*, **U. Anwar***, A. Ahmed, and A. Aghasi. Learning to solve differential equations across initial conditions. In *ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations*, 2020. URL arxiv.org/abs/2003.12159

Work Experience

Senior Machine Learing Engineer February 2020 – March 2022
Scientific Computing Department
NetSol Technologies, Lahore

Research Assistant & Graduate Student Fellow July 2019 – June 2020
Center of Artificial Intelligence and Computational Science
Information Technology University, Lahore

Research Intern July – September 2018
Internet of Things Laboratory
Khwarizmi Institute of Computer Science, Lahore

Junior Data Scientist June – August 2017
ADDO AI, Lahore

Teaching Experience

Teaching Assistant - Discrete Mathematics September 2019 – January 2020
Department Of Computer Science
Information Technology University, Lahore

(Selected) Talks

Foundational Challenges in LLM Alignment and Safety - SERI MATS, ACM Summer School on Responsible & Safe AI, GovAI, Chalmers AI Ethics Seminar and others.

Reward Modelling for AGI Safety - Future of Life Seminar on AI Safety, October 2022.

Professional Activities & Services

Lead Organizer: NeurIPS Workshop on Socially Responsible Language Modelling Research (SoLaR), 2023 & 2024

Peer Reviewer: ICML 2023, NeurIPS 2023, ICLR 2025

Grant Reviewer: Vitalik Buterin Fellowship for AI Existential Safety (2023)

Mentor: GradAppLab (2022 - Present)

Mentoring & Supervisions

Karim Abdul Sadek *for* Mitigating Goal Misgeneralization via Minimax Regret at University of Cambridge, UK (resulted in ICLR 2025 submission).

Thomas Bush *for* Interpreting Emergent Planning in Model-Free Reinforcement Learning at University of Cambridge, UK (resulted in ICLR 2025 submission).

Miguel Lara *for* Learning to Forget Using Diffusion Hypernetworks at University of Cambridge, UK (resulted in NeurIPS workshop paper).

Thomas Coste *for* Conservative Agency For Safe Optimization of Learned Reward Models at University of Cambridge, UK (resulted in ICLR 2024 paper).

Yawen Duan *for* Red Teaming (Learned) Reward Models at University of Cambridge, UK.

Jason Brown *for* Likert Scale Feedback to Improve Robustness of Learning From Human Preferences at University of Cambridge, UK.

Abdul Rehman & Arslan Malik *for* Privacy Preserving Recommendation System at ITU, Pakistan.

Volunteer Work

Managing Director & Co-Founder Spectra Magazine April 2017 – May 2020
Spectra Magazine is a student-powered online magazine aiming to enhance public understanding of science and shape the narrative of science journalism in Pakistan. Under my leadership, we published more than 215 articles and mentored more than 50 high school and undergraduate students in science writing, editing and design. Read more about us at *www.spectramagazine.org/about*.

Non-Degree Studies

Cooperative AI Summer School *(July 2023)*
Eastern European Machine Learning School *(July 2021)*

Skills

• Python (Numpy, Scipy, Matplotlib) • Jax • Pytorch • Tensorflow • C • SQL • NoSQL