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: 2nd, 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

Government College University, Lahore August 2013 – May 2015
Associate's Degree Pre-Engineering (Grade: A+)

Awards & Honours

2022 Open Phil AI Fellowship by Open Philanthrophy Foundation. 🔗

2022 Vitalik Buterin PhD Fellowship by Future Of Life Institute. 🔗

Free Registration Award at virtual MLSS 2021 Taipei.

Graduate Student Fellowship, ITU, Lahore.

Merit Scholarship, ITU, Lahore.

Preprints

U. Anwar, Y. Choi, D. Chen, F. Tramer, H. He, A. Kasirzadeh, J. Foerster, D. Krueger, et al. Foundational challenges in LLM alignment and safety, 2024

Conference & Journal Publications

** denotes equal contribution*

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

Workshop Publications

U. Anwar, J. Wan, D. Krueger, and J. Foerster. Noisy ZSC: Breaking the common knowledge assumption in zero-shot coordination games. In *NeurIPS 2023 Workshop on Open Ended Learning*, 2023

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

Talks

Foundational Challenges in LLM Alignment and Safety - SERI MATS, August 2023.

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

Selected Ongoing Projects

Adversarial Robustness of In-Context Learning
Usman Anwar, Spencer Frei, Johannes Von Oswold, David Krueger

I am leading a project which aims at understanding adversarial robustness of in-context learning in various settings, and designing methods to make in-context learning adversarially robust.

Unifying Goal Misgeneralization in RL and Non-Identifiability of Reward Function in Inverse RL
Usman Anwar, Matthew Farrugia-Roberts, David Krueger

I am leading a project which aims to formalize goal misgeneralization as a special case of underspecification and misspecification in task specification. This will show that outer alignment and inner alignment problems perhaps arise from similar limitations in the training process.

Professional Activities & Services

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

Peer Reviewer: ICML 2023, NeurIPS 2023

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

Mentor: GradAppLab (2022 - Present)

Mentoring & Supervisions

Thomas Coste *for* Conservative Agency For Safe Optimization of Learned Reward Models at University of Cambridge, UK.

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) • Pytorch • Tensorflow • C • SQL • NoSQL