ROBERTA RAILEANU

11-12 Canal Reach, London, UK
raileanu@meta.com ⋄ rraileanu.github.io ⋄ @robertarail

INTERESTS

AI Scientist, Agents, Tool Use, LLMs, Reinforcement Learning, Open-Ended Learning

CURRENT POSITION

Research Scientist at Meta, GenAI, Llama Research Agents Team Oct 2021 - Present

EDUCATION

New York University, NY, USA Sep 2016 - Sep 2021
PhD in Computer Science
Thesis: Towards More General and Adaptive Reinforcement Learning Agents
Advisor: Rob Fergus

Princeton University, NJ, USA Sep 2012 - June 2016
A.B. in Astrophysical Sciences, *magna cum laude*
Certificates (Minors): Statistics and Machine Learning, Applications of Computing
Thesis: Clustering Redshift Estimation for the Hyper Suprime-Cam Survey
Advisor: Michael Strauss

PUBLICATIONS

Llama Team, The Llama 3 Herd of Models, *arXiv*, 2024. Tool Use Lead.

Deepak Nathani, Lovish Madaan, Nicholas Roberts, Nikolay Bashlykov, Ajay Menon, Vincent Moens, Amar Budhiraja, Despoina Magka, Vladislav Vorotilov, Gaurav Chaurasia, Dieuwke Hupkes, Ricardo Silveira Cabral, Tatiana Shavrina, Jakob Foerster, Yoram Bachrach, William Yang Wang, **Roberta Raileanu**. MLGym: A New Framework and Benchmark for Advancing AI Research Agents. *arXiv*, 2025.

Martin Klissarov, Mikael Henaff, **Roberta Raileanu**, Shagun Sodhani, Pascal Vincent, Amy Zhang, Pierre-Luc Bacon, Doina Precup, Marlos C. Machado, Pierluca D'Oro. MaestroMotif: Skill Design from Artificial Intelligence Feedback. *ICLR*, 2025 (**oral**).

Samvelyan M, Raparthy SC, Lupu A, Hambro E, Markosyan AH, Bhatt M, Mao Y, Jiang M, Parker-Holder J, Foerster J, Rocktäschel T, **Raileanu R**, Rainbow teaming: Open-ended generation of diverse adversarial prompts, *NeurIPS*, 2024.

Havrilla A, Du Y, Raparthy SC, Nalmpantis C, Dwivedi-Yu J, Zhuravinskyi M, Hambro E, Sukhbaatar S, **Raileanu R**, Glore: When, where, and how to improve llm reasoning via global and local refinements, *ICML*, 2024.

Havrilla A, Du Y, Raparthy SC, Nalmpantis C, Dwivedi-Yu J, Zhuravinskyi M, Hambro E, **Raileanu R**, Teaching large language models to reason with reinforcement learning, *arXiv*, 2024.

Kirk R, Mediratta I, Nalmpantis C, Luketina J, Hambro E, Grefenstette E, **Raileanu R**, Understanding the Effect of RLHF on LLM Generalisation and Diversity, *ICLR*, 2024.

Raparthy SC, Hambro E, Kirk R, Henaff M, **Raileanu R**, Generalization to new sequential decision making tasks with in-context learning, *ICML*, 2024.

Earle S, Kokkinos F, Nie Y, Togelius J, **Raileanu R**, Dreamcraft: Text-guided generation of functional 3D environments in Minecraft, *FDG*, 2024, **(best paper award)**.

Kirk R, Mediratta I, Nalmpantis C, Luketina J, Hambro E, Grefenstette E, **Raileanu R**, The Effect of Reinforcement Learning from Human Feedback on the Generalisation and Diversity of Large Language Models, *ICML Workshop*, 2023.

Kaddour J, Harris J, Mozes M, Bradley H, **Raileanu R**, McHardy R, Challenges and Applications of Large Language Models, *arXiv*, 2023.

Schick T, Dwivedi J, Dessi R, **Raileanu R**, Lomeli M, Zettlemoyer L, Canceda N, Scialom T,Toolformer: Language Models Can Teach Themselves to Use Tools, *arXiv*, 2023.

Mialon et al., Augmented Language Models: A Survey, *arXiv*, 2023.

Chen Y, Marchisio K, **Raileanu R**, Adelani DI, Stenetor P, Riedel S, Improving Language Plasticity via Pretraining with Active Forgetting, *arXiv*, 2023.

Jiang Y, Kolter Z, **Raileanu R**, On the Importance of Exploration for Generalization in Reinforcement Learning, *under review*, 2023.

Eimer T, Lindauer M, **Raileanu R**, Hyperparameters in Reinforcement Learning and How to Tune Them, *ICML*, 2023.

Henaff M, Jiang M, **Raileanu R**, A Study of Global and Episodic Bonuses for Exploration in Contextual MDPs, *ICML*, 2023 **(oral)**.

Gaya JB, Doan T, Caccia L, Soulier L, Denoyer L, **Raileanu R**, Building a Subspace of Policies for Scalable Continual Learning, *ICLR*, 2023 (**spotlight, top-25%**).

Samvelyan M, Khan A, Dennis M, Jiang M, Parker-Holder J, Foerster J, **Raileanu R**, Rocktäschel T, MAESTRO: Open-Ended Environment Design for Multi-Agent Reinforcement Learning, *ICLR*, 2023.

Henaff M, Jiang M, **Raileanu R**, Integrating Episodic and Global Novelty Bonuses for Efficient Exploration, *under review*, 2023.

Henaff M, **Raileanu R**, Jiang M, Rocktäschel T, Exploration via Elliptical Episodic Bonuses, *NeurIPS*, 2022.

Mu J, Zhong V, **Raileanu R**, Jiang M, Goodman N, Rocktäschel T, Grefenstette E, Improving Intrinsic Exploration with Language Abstractions, *NeurIPS*, 2022.

Hambro E, **Raileanu R**, Rothermel D, Mella V, Rocktäschel T, Kuttler H, Murray N, Dungeons and Data: A Large-Scale NetHack Dataset, *NeurIPS*, 2022.

Open Ended Learning Team, Stooke A, Mahajan A, Barros C, Deck D, Bauer J, Sygnowski J, Trebacz M, Jaderberg M, Mathieu M, McAleese N, Bradley-Schmieg N, Wong N, Porcel N, **Raileanu R**, Hughes-Fitt S, Dalibard V, Czarnecki W, Open-Ended Learning Leads to Generally Capable Agents, *arXiv*, 2021.

Raileanu R, Fergus R, Decoupling Value and Policy for Generalization in Reinforcement Learning, *ICML*, 2021 (**oral**).

Raileanu R, Goldstein M, Yarats D, Kostrikov I, Fergus R, Automatic Data Augmentation for Generalization in Deep Reinforcement Learning, *NeurIPS*, 2021 and *Inductive Biases, Invariances, and Generalization in Reinforcement Learning Workshop*, *ICML*, 2020 (**oral**).

Campero A, **Raileanu R**, Heinrich K, Tenenbaum J, Rocktäschel T, Grefenstette E, Learning with AMIGo: Adversarially Motivated Intrinsic Goals, *ICLR*, 2021.

Raileanu R, Goldstein M, Szlam A, Fergus R, Fast Adaptation to New Environments via Policy-Dynamics Value Functions, *ICML* 2020 and *Beyond "Tabula Rasa" in Reinforcement Learning Workshop*, *ICLR*, 2020 (**oral**).

Raileanu R, Rocktäschel T, RIDE: Rewarding Impact-Driven Exploration for Procedurally-Generated Environments, *ICLR*, 2020.

Heinrich K, Nardelli N, Miller A, **Raileanu R**, Selvatici M, Grefenstette E, Rocktäschel T, The NetHack Learning Environment, *NeurIPS*, 2020.

Resnick C*, **Raileanu R***, Kapoor S, Peysakhovich A, Cho K, Bruna J, Backplay: "Man Muss Immer Umkehren", *Reinforcement Learning in Games Workshop*, *AAAI*, 2019.

Raileanu R, Denton E, Szlam A, Fergus R, Modeling Others using Oneself in Multi-Agent Reinforcement Learning, *ICML*, 2018.

Raileanu R, Szlam A, Fergus R, Modeling Other Agents' Hidden States in Deep Reinforcement Learning, *Emergent Communication Workshop*, *NeurIPS*, 2017.

Kim CK, Ostriker EC, **Raileanu R**, Superbubbles in the Multiphase ISM and the Loading of Galactic Winds, *The Astrophysical Journal*, 2016.

RESEARCH EXPERIENCE

DeepMind, London, UK Jan 2021 - Jun 2021
Research Intern

Researched unsupervised environment design methods for generalization in 3D environments.
Advisor: Max Jaderberg

Facebook AI Research, London, UK June - Sep 2019
Research Intern

Developed a new algorithm for exploration in sparse reward procedurally-generated environments.
Advisor: Tim Rocktäschel

Microsoft Research, Cambridge, UK June - Aug 2018
Research Intern

Researched methods for zero-shot and few-shot generalization in multi-agent settings.
Advisors: Katja Hofmann, Sam Devlin

Facebook AI Research, New York, USA June - Aug 2017
Research Intern

Researched methods for modeling other agents in semi-cooperative reinforcement learning settings.
Advisor: Arthur Szlam

Princeton University, Princeton, USA June - Aug 2015
Undergraduate Researcher

Developed 3D hydrodynamical simulations of supernovae in the multiphase interstellar medium.
Advisors: Eve Ostriker, Chang-Goo Kim

Princeton University, Princeton, USA Feb - May 2015
Undergraduate Researcher

Implemented and evaluated machine learning techniques for the prediction of stellar rotation periods.
Advisor: Timothy Morton

ETH, Zürich, Switzerland Jun - Aug 2014
Research Intern

Created Monte Carlo simulations for exoplanet detection with the James Webb Space Telescope.
Advisor: Michael Meyer

Max Planck for Extraterrestrial Physics, Garching, Germany Jun - Aug 2013
Research Intern

Developed N-Body simulations and theoretical models of the Milky Way Galaxy.
Advisor: Ortwin Gerhard

HONORS & AWARDS

Rising Stars in EECS	2020
Sigma Xi: Scientific Research Honor Society	2016
Bell Burnell Award for Early Career Female Physicist	2013
Bronze Medal at the International Physics Olympiad	2012
Silver Medal at the International Physics Olympiad	2011
Gold Medal at the International Astrophysics Olympiad	2011
Silver Medal at Tuymaada International Olympiad in Physics	2010

INVITED TALKS AND PANELS

Air Street London.AI Meetup	Nov 2024
Romanian AI Days	Sep 2024
EEA NLP for Space Science Workshop	Sep 2024
RLC Training Agents with Foundation Models	Aug 2024
Oxford Workshop on RobustLLM	July 2024
ICML Workshop on AutoRL	July 2024
ICLR Workshop on LLM Agents	May 2024
RLC Workshop on Training Agents with Foundation Models	May 2024
Center for Theoretical Neuroscience, Columbia University	April 2024
M2L Summer School: Introduction to Reinforcement Learning	Aug 2023
ICLR Generalization in RL Workshop	Apr 2023
CarperAI: Augmenting LLMs with Tools	Mar 2023
Microsoft Research: Augmenting LLMs with Tools	Mar 2023
NYU: Augmenting LLMs with Tools	Mar 2023
Neural MMO Open-Endedness Panel	Oct 2022
AI and Games Summer School	Aug 2022
Imperial ICARL Seminar	May 2022
Microsoft Research Summit	Aug 2021
Princeton Intelligent Robot Motion Lab	Mar 2021
Berkeley Rising Stars EECS	Nov 2020
NYU Game Innovation Lab	Jul 2020

MENTORING AND MANAGING EXPERIENCE

Deepak Nathani, Intern, Meta - *advancing AI research with AI*	2024
Alexander Havrilla, Intern, FAIR - *improve LLM reasoning via RL*	2022
Dheeraj Mekala, Intern, FAIR - *LLM generalization to new tools*	2022
Carlos Gemmell, Intern, FAIR - *teach LLMs to use SQL*	2022
Yuqing Du, Visiting Researcher, FAIR - *improve LLM generation via RL*	2022
Hao Lio, Intern, FAIR - *LLMs for sequential decision making*	2022
Shehzaad Dhuliawala, Intern, FAIR - *improving LLM hallucinations via self-verification*	2022
Martin Klissarov, Intern, FAIR - *RLHF for playing hard exploration games*	2022
Alon Albalak, Intern, MSR - *curricula for more efficient LLM training*	2022
Yiding Jiang, Intern, FAIR - *exploration for generalization in RL*	2022
Rob Kirk, Intern, FAIR - *finetuning LLMs with RL and SL*	2022

Theresa Eimer, Intern, FAIR - *training web agents with RL* 2022
Sam Earle, Intern, FAIR - *text-guided world generation in MineCraft* 2022
Jean-Baptiste Gaya, PhD Student, FAIR - *continual reinforcement learning* 2022
Sharath Chandra, AI Resident, FAIR - *few-shot learning of new behaviors* 2022
Ishita Mediratta, AI Resident, FAIR - *generalization in sequential decision making* 2022
Minqi Jiang, PhD Student, FAIR - *open-ended learning* 2022
Mikayel Samvelyan, PhD Student, FAIR - *open-ended learning for MARL* 2022
Yingchen Xu, PhD Student, FAIR - *self-supervised reinforcement learning* 2022
Jesse Mu, Intern, FAIR - *language for exploration* 2021
Aaron Roth, PhD Student, UMD (now US Naval Research Lab) - *representation learning* 2020
Chang Ye, MS Student, NYU (now Google) - *adaptation to new environments* 2020
Srikar Yellapragada, MS Student, NYU (now Stony Brook) - *RL for translation* 2019
Chandra Konkimalla, MS Student, NYU (now Amazon) - *learning from demonstrations* 2019
Zeping Zhan, MS Student, NYU (now Kooick) - *multi-agent learning in social dilemmas* 2019

PHD SUPERVISION

Nathan Herr, UCL - *planning and reasoning in LLMs* 2024 - Present
Alisia Lupidi, Oxford-Meta - *interactive LLMs* 2023 - Present

PHD COMMITTEE

Eduardo Pignatelli, UCL 2024
Mathieu Rita, INRIA & ENS 2024
Laetita Teodorescu, INRIA 2023

REVIEWING EXPERIENCE

2023: ICLR, ICML, NeurIPS
2022: ICLR, ICML, NeurIPS, EWRL, ICLR GMS Workshop
2021: ICLR, ICML, NeurIPS
2020: ICLR, ICML, NeurIPS, UAI, ICML LAOW Workshop, IEEE
2019: ICLR, ICML, NeurIPS, ICML I3 Workshop
2018: ICLR, ICML, NeurIPS

ORGANIZING EXPERIENCE

Reward Free RL Workshop at RLC 2024
Generative Models for Decision Making at ICLR 2024
Socially Responsible Language Modelling Research (SoLaR) Workshop at NeurIPS 2023
Agent Learning in Open-Endedness (ALOE) Workshop at ICLR 2022 and NeurIPS 2023
Unsupervised Reinforcement Learning (URL) Workshop at ICML 2021

TEACHING EXPERIENCE

African Master's of Machine Intelligence (AMMI), Kigali, Rwanda – NLP March 2019
Princeton McGraw Center, New Jersey, USA – Math, Physics 2015 - 2016

RELEVANT SKILLS

PyTorch, JAX, Tensorflow, Python, Java, Matlab, R, C++, OCaml