Karolis Jucys

(prev. Karolis Ramanauskas)

K@rol.is
ka.rol.is

Experience

10/2022 – present **PhD Student**, *University of Bath*, UK
Interpretability of vision-based agents.

10/2018 – 09/2022 **Independent AI Researcher**, *Self-employed*, UK
- Organizer of MineRL 2022 and 2021 competitions.
- Finalist at The NetHack Challenge, MineRL 2020, MineRL 2019, Unity Obstacle Tower Challenge.

05/2019 – 09/2019 **Research Engineer**, *Neurotechnology*, Vilnius, Lithuania, Part-time
- Applying deep learning techniques to financial markets; researching autonomous driving and navigation.

01/2015 – 06/2017 **Quantitative Analyst**, *Praxis Partners LLP*, London, UK

07/2011 – 01/2015 **Quantitative Analyst**, *Blue Edge Capital Partners LLP*, London, UK

01/2010 – 06/2011 **Analyst**, *Reditus Partners*, Vilnius, Lithuania

Education

2009 **MSc, Mathematics**, *Vilnius University*

2007 **BSc, Computer Science**, *Vilnius University*

Certifications

2019 **Reinforcement Learning Specialization**, *Coursera – University of Alberta*

2018 **Deep Learning Specialization**, *Coursera*

2013 **Machine Learning**, *Coursera – Stanford University*

Volunteering

2015 – Present **Mentor**, *LT Big Brother*

2022 – Present **Member**, *Giving What We Can*

Languages

English, Russian, Lithuanian

Achievements

OpenAI Safety Fellowship – finalist.
Anthropic AI Safety Fellowship – finalist (top 3–4%).
Google Code Jam 2017 – #1,280 out of 60,000+.
Topcoder SRM 614 Division 2 – 1st place (out of 1,173).
ModelOff 2015, 2016 and 2018 – finalist (top 16 out of ~4,000).
HIQORA High IQ World Championships 2016 – finalist (top 12 out of 2,000+).
Rubik's Cube – 3min blindfolded solve, 18s regular solve.
Savoy Cup international team dance competition – 1st place in 2016 and 2017.
Opening act for the season 12 premiere of BBC's Dragons' Den – dancer (2.63M TV audience).
Runescape Hardcore Ironman – rank 3 (out of 230,000+).
Minecraft Speedrunning – Any% Glitchless in 36m 17s.

Publications

2026 **MultiSTEVE-1s: A Model Zoo and Interpretability Suite for Instruction-Following Vision Agents**
K Jucys, G Adamopoulos, Ö Şimşek
(Spotlight – top ~3% of submissions) ICML 2026 Mechanistic Interpretability Workshop

2024 **Interpretability in Action: Exploratory Analysis of VPT, a Minecraft Agent**
K Jucys, G Adamopoulos, M Hamidi, S Milani, MR Samsami, A Zholus, S Joseph, ...
ICML 2024 Mechanistic Interpretability Workshop

2023 **Colour versus Shape Goal Misgeneralization in Reinforcement Learning: A Case Study**
K Ramanauskas, Ö Şimşek
NeurIPS 2023 ATTRIB: Workshop on Attributing Model Behavior at Scale

2023 **BEDD: The MineRL BASALT Evaluation and Demonstrations Dataset for Training and Benchmarking Agents that Solve Fuzzy Tasks**
S Milani, A Kanervisto, **K Ramanauskas**, SV Schulhoff, B Houghton, J Shen, ...
(Oral – top ~1% of submissions) NeurIPS 2023 Track on Datasets and Benchmarks

2023 **Towards Solving Fuzzy Tasks with Human Feedback: A Retrospective of the MineRL BASALT 2022 Competition**
S Milani, A Kanervisto, **K Ramanauskas**, S Schulhoff, B Houghton, S Mohanty, ...
NeurIPS 2022 Competitions Track, 171-188

2022 **Insights from the NeurIPS 2021 NetHack challenge**
E Hambro, S Mohanty, D Babaev, M Byeon, D Chakraborty, E Grefenstette, ...
NeurIPS 2021 Competitions and Demonstrations Track, 41-52

2022 **Minerl diamond 2021 competition: Overview, results, and lessons learned**
A Kanervisto, S Milani, **K Ramanauskas**, N Topin, Z Lin, J Li, J Shi, D Ye, Q Fu, ...
NeurIPS 2021 Competitions and Demonstrations Track, 13-28

2021 **Towards robust and domain agnostic reinforcement learning competitions: MineRL 2020**
WH Guss, S Milani, N Topin, B Houghton, S Mohanty, A Melnik, A Harter, B Buschmaas, ...
NeurIPS 2020 Competition and Demonstration Track, 233-252

Talks / Media

2025 Cohere. "Interpretability in Action Exploratory Analysis of VPT, a Minecraft Agent".

2024 ScienceNewsExplores. Interviewed for the article "How to design artificial intelligence that acts nice — and only nice".

2023 Lithuanian Embassy in London. "AI Safety / Risk".

2021 NeurIPS. "Competition Track" workshop. "Diamond: A MineRL Competition on Training Sample-Efficient Agents".

2021 NeurIPS. "Competition Track" workshop. Finalist presentation at "The NetHack Challenge".

2021 Microsoft Research Summit. "Minecraft as a research platform for reinforcement learning" workshop. "Demo: MineRL Diamond & Basalt".

2021 Ignition Hacks. "Learning from Human Demonstrations".

2019 NeurIPS. "Competition Track" workshop. Finalist presentation at "The MineRL competition".

Reviewing

2026 Reviewing for ICML 2026 Mechanistic Interpretability Workshop.

2022–2026 Reviewing for NeurIPS Competition Track.

2025 Reviewing for NeurIPS 2025 Mechanistic Interpretability Workshop.

2022, 2025 Judging for the Generative Design in Minecraft Competition.

2023 Reviewing the Distributional Reinforcement Learning book.