

# Grgur Kovač

PHD STUDENT AT FLOWERS TEAM (INRIA)

✉ kovac.grgur@gmail.com 🏠 Website 🔗 LinkedIn ✖ X (Twitter) 🐦 Bluesky 🎓 Google Scholar

## Work Experience

### PhD student at Flowers team - INRIA

Bordeaux, France

PHD STUDENT, SUPERVISED BY PIERRE-YVES OUDEYER AND PETER FORD DOMINEY

Mar 2022 - Jul 2025

- research on how we can leverage human sciences (psychology, cultural evolution) to better understand, evaluate, and build RL agents and Large Language Models ([thesis](#), to defend 5th Nov 2025)
- [Recursive Training Loops in LLMs](#): adapting cultural evolution methodology to study which properties of internet data might increase or mitigate degradation of LLM generated text following iterative fine-tuning ([paper](#), EMNLP 2025 Oral)
- [StickToYourRole](#): adapting psychological methodology to AI to study stability of values expressed by LLM-simulated populations ([paper](#)); and further extending the methodology and maintaining the [StickToYourRole Leaderboard](#) (>24.6k visits)
- [LLM's as superpositions of cultural perspectives](#): a position paper discussing and demonstrating LLMs sensitivity to trivial context changes and the implications of that for evaluation and understanding LLM simulated behavior ([paper](#), cited >70 times)
- [SocialAI](#): following developmental theories of M. Tomasello and J. Bruner to outline core socio-cognitive abilities and concepts for AI, and constructing a tool (procedural environment generator) to foster research of those concepts in RL- and LLM- based agents
- Second Author: following cultural evolution uncovering biases and attractors in stories iteratively generated by LLMs ([paper](#), ICLR 2025)
- Second Author: presenting an architecture leveraging Learning Progress estimates and LLMs to generate Craftax environments for an RL agent ([preprint](#))
- Minor project: creating first version of [LLM4Humanties](#) tool for qualitative analysis with LLMs (taken on by another student)
- *Technologies used*: Python, PyTorch, Transformers, Unsloth, vLLM, SLURM, scikit-learn, Jupyter

### Research engineer at Flowers team - INRIA

Bordeaux, France

RESEARCH ENGINEER

Nov 2019 - Jan 2022

- research in the field of Deep Reinforcement Learning, and at the intersection of AI and cognitive science
- [GRIMGEP](#): creating an architecture augmenting novelty based exploration of RL agents with Absolute Learning Progress
- [SocialAI](#): following developmental theories of M. Tomasello and J. Bruner to outline core socio-cognitive abilities and concepts for AI, and constructing a tool (procedural environment generator) to foster research of those concepts in RL- and LLM- based agents
- *Technologies used*: Python, PyTorch, SLURM, scikit-learn

### Research engineer at Microblink / Photomath

Zagreb, Croatia

STUDENT JOB

Jul 2017 - Sep 2019

- collaboratively developed production deep learning models for OCR technology and other computer vision tasks
- created a simple text-based receipt classifier which in production classified over 200 million receipts in one year
- creating a method to train binary neural networks with small losses in accuracy
- collaboratively participating in the efficiency MicroNet challenge (at NeurIPS 2019) - 6th from 19 participating teams
- *Technologies used*: Python, Tensorflow, scikit-learn, Docker, Kubernetes

### Student assistant

Zagreb, Croatia

FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING, UNIVERSITY OF ZAGREB

Mar 2019 - Jun 2019

- assisting in the deep learning course by evaluating students' homework and exams

## Education

### PhD in Computer Science

Bordeaux, France

FLOWERS TEAM, INRIA

Mar 2023 - Jul 2025

- [Thesis](#): Building, evaluating and understanding socio-cultural AI: leveraging concepts and methods from human sciences
- adapting psychology and cultural evolution theories and methodology to better understand, evaluate, and build RL agents and Large Language Models

### Master of Computer Science

Zagreb, Croatia

FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING, UNIVERSITY OF ZAGREB

2017/18 - 2018/19

- master's thesis on the topic on Multiple Object Tracking ([source code](#))
- text analysis course project on the topic of information retrieval ([project report](#))

- bachelor's thesis: A framework for training feed-forward fully connected neural networks ([source code](#))

## Publications

---

### PEER-REVIEWED

- **Grgur Kovač\***, Jérémy Perez\*, Rémy Portelas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2025. Recursive Training Loops in LLMs: How training data properties modulate distribution shift in generated data? In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025, Oral)
- Jérémy Perez, **Grgur Kovač**, Corentin Léger, Cédric Colas, Gaia Molinaro, Maxime Derex, Pierre-Yves Oudeyer, and Clément Moulin-Frier (2025). 'When LLMs Play the Telephone Game: Cultural Attractors as Conceptual Tools to Evaluate LLMs in Multi-turn Settings'. In: *The Thirteenth International Conference on Learning Representations*
- **Grgur Kovač**, Rémy Portelas, Masataka Sawayama, Peter Ford Dominey, and Pierre-Yves Oudeyer (2024b). 'Stick to your Role! Stability of Personal Values Expressed in Large Language Models'. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 46
- **Grgur Kovač**, Rémy Portelas, Masataka Sawayama, Peter Ford Dominey, and Pierre-Yves Oudeyer (Aug. 2024a). 'Stick to your role! Stability of personal values expressed in large language models'. In: *PLOS ONE* 19.8
- **Grgur Kovač**, Rémy Portelas, Peter Ford Dominey, and Pierre-Yves Oudeyer (2024). 'The SocialAI school: a framework leveraging developmental psychology toward artificial socio-cultural agents'. In: *Frontiers in Neurorobotics* Volume 18 - 2024
- **Grgur Kovač**, Adrien Laversanne-Finot, and Pierre-Yves Oudeyer (2022). 'Grimgep: learning progress for robust goal sampling in visual deep reinforcement learning'. In: *IEEE Transactions on Cognitive and Developmental Systems* 15.3

### PREPRINTS

- **Grgur Kovač**, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer (2023). 'Large language models as superpositions of cultural perspectives'. In: *arXiv preprint arXiv:2307.07870*

### WORKSHOPS

- **Grgur Kovač\***, Rémy Portelas\*, Katja Hofmann, and Pierre-Yves Oudeyer (June 2021). 'SocialAI 0.1: Towards a Benchmark to Stimulate Research on Socio-Cognitive Abilities in Deep Reinforcement Learning Agents'. In: NAACL. Accepted at NAACL ViGIL Workshop 2021. Mexico City, Mexico (Spotlight)
- **Grgur Kovač**, Rémy Portelas, Peter Ford Dominey, and Pierre-Yves Oudeyer (July 2023). 'The SocialAI School: Insights from Developmental Psychology Towards Artificial Socio-Cultural Agents'. In: TOM 2023 -First Workshop on Theory of Mind in Communicating Agents - ICML 2023 Workshop. Honolulu (Hawaii), United States
- Guillaume Pourcel, Thomas Carta, **Grgur Kovač**, and Pierre-Yves Oudeyer (2024). 'Autotelic LLM-based exploration for goal-conditioned RL'. In: Intrinsically Motivated Open-ended Learning Workshop at NeurIPS 2024

\*equal contribution

### TOOLS AND SOFTWARE

- [StickToYourRole leaderboard](#) - compares LLMs based on undesired sensitivity to context change on the task of simulating populations
- [LLM4Humanities](#) - a tool for qualitative analysis with LLMs

## Other interests and skills

---

### LANGUAGE SKILLS

- **Croatian**: native
- **English**: fluent
- **French**: conversational

### HOBBIES

- Maintaining old/vintage bicycles