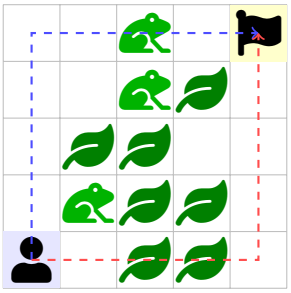
Teaching Agents to Follow the Rules: Normative Reinforcement Learning

Advisor: **Matthias Grilz, Liam Plank**

Motivation

Autonomous agents should not only be good at their task, they should also follow rules: safety regulations, codes of conduct, or ethical norms. But norms can conflict and come with priorities, which standard reinforcement learning (RL), with its single reward signal, struggles to express.

Building on a framework recently developed in our group, this topic explores how to best bring ethical and normative concerns into RL, tested on gridworld benchmarks.



We have to reach the goal. **Which route should we take?**

To reach the goal, the agent must either step on a frog or on two plants. This is where rule priorities come into play: You need to teach the agent how to decide which rule it may violate and which one it must respect.

Goals and Tasks

- Become familiar with normative reinforcement learning and the existing framework
- Explore a direction of your choice
  - Multi-objective RL
  - Constrained RL
  - Game settings with other agents
- Implement and evaluate the approach on our benchmarks

Literature

- E. A. Neufeld, A. Ciabattoni, and R. F. Tulcan
  Combining MORL with Restraining Bolts to Learn Normative Behaviour
  https://doi.org/10.24963/ijcai.2025/514
- C. F. Hayes et al.
  A Practical Guide to MORL & Planning
  2022
  https://doi.org/10.1007/s10458-022-09552-y

Courses & Deliverables

☑ **Master Project**
Project code
Report
Presentation

– OR –

☑ **Master's Thesis**
Initial presentation
Project code
Thesis (60+ pages)
Final presentation

Recommended if you're studying

☑ CS ☑ ICE ☑ SEM

Prerequisites

- Interest in RL, AI safety, and AI ethics
- Programming (Python)

Advisor Contact

trustedai.isec@mlist.tugraz.at