

Subrat Prasad Panda

Reinforcement Learning Researcher

Neurosymbolic RL · Planning · LLM Embodied Agents · Safe AI

✉ subratpr001@e.ntu.edu.sg
☎ +65 ****3021 · Singapore
🌐 [linkedin.com/in/subratpp](https://www.linkedin.com/in/subratpp)
🐙 github.com/subratpp
🌐 subratpp.github.io
👤 Google Scholar

RL researcher (Ph.D. candidate, NTU Singapore) focused on **neurosymbolic RL**: agents that are **sample-efficient, robust, interpretable and verifiably safe**. 5+ years with **deep RL** (PPO, SAC, distributional and hierarchical RL), **MCTS** planning, **LLM-based planning for embodied agents** and **RL fine-tuning of LLMs** (GRPO); hands-on robotics with **ROS 2** on Jetson. Former ML Engineer at InFocus (contract project for **Google X**); first-author papers at **JMLR, ECML-PKDD** and **ECAI**.

Technical Skills

Reinforcement Learning	PPO, SAC, A2C, QR-DQN / distributional RL, option-critic & hierarchical RL, goal-conditioned RL, reward shaping, sparse-reward & long-horizon tasks, robust & safe RL, neurosymbolic & programmatic policies, world models/MDP latent models (WAE-MDP)
Planning & Search	Monte Carlo Tree Search (MCTS), belief-space planning, D* / A* replanning, symbolic task planners, LLM-based planning for embodied agents
LLMs & Deep Learning	RL fine-tuning of LLMs (GRPO, RLHF, DPO), transformers from scratch, LLM steering for safe embodied planning, LLM-generated policies, Decision Transformers, mechanistic interpretability, MoE-style Top- <i>k</i> routing, variational encoders, differentiable decision trees, computer vision
Verification	Formal verification of RL controllers, reachability analysis, abstract interpretation, LP/MILP (Gurobi)
Tools	Python, PyTorch, Hugging Face (Transformers, TRL), Stable-Baselines3, CleanRL, Gymnasium / OpenAI Gym, MuJoCo, MiniGrid, ROS 2, SLAM & localisation (EKF, particle filter), NVIDIA Jetson AGX, MATLAB, Arduino, $\LaTeX$

Education

2023 – 2027 expected	Ph.D., Computer Science · Nanyang Technological University (NTU) Thesis: <i>Neurosymbolic Techniques for Safe and Efficient RL</i> . Advisors: Arvind Easwaran (NTU), Blaise Genest (CNRS).	Singapore
2019 – 2021	M.Tech., Computer Science · Indian Statistical Institute (ISI) Thesis: <i>User allocation & service placement in edge computing</i> . Advisor: Ansuman Banerjee. CPI 82.1%; top-5 thesis nomination.	Kolkata, India
2012 – 2016	B.Tech., Electronics & Instrumentation · NIT Silchar Thesis: <i>Real-time DC motor position control with PID and fuzzy logic</i> . Advisor: Munmun Khanra.	Silchar, India

Experience

2023 – Present	Ph.D. Researcher, Safe & Efficient RL <i>Nanyang Technological University (NTU) & CNRS@CREATE</i> ▶ Led 5 end-to-end RL research systems (hierarchical RL, MCTS planning, distributional RL, formal verification); 2 open-sourced.	Singapore
2022 – 2023	Research Associate <i>CNRS@CREATE</i> ▶ Built a deep reinforcement learning controller for fine-grained, real-time energy management of a grid-connected microgrid. ▶ Mitigated forecast uncertainty in real-time operation (<i>Sustainable Energy, Grids and Networks</i>, Elsevier 2024); extended to interpretable programmatic RL policies (AI4X 2025).	Singapore
2021 – 2022	Machine Learning Engineer <i>InFocus Innovations, on attachment to Google X (USA)</i> ▶ Developed reinforcement learning agents for the Abstraction and Reasoning Corpus (ARC), a benchmark for abstract reasoning and program synthesis, as part of a Google X project.	Pune, India
2021 – 2021	Research Engineer, Autonomous Vehicles <i>TiHAN, IIT Hyderabad</i> ▶ Built drive-by-wire braking and throttle control for an autonomous test vehicle on Arduino. ▶ Deployed a lane-detection deep learning model on the NVIDIA Jetson AGX edge platform.	Hyderabad, India

Selected Projects

ECML-PKDD
2026
published

Neurosymbolic Hierarchical RL with MCTS Planning

 code

STACK PyTorch · MuJoCo AntMaze · Point-Maze · MCTS · D* · goal-conditioned RL

- ▶ Two-level agent: a **symbolic planner** with an online-updated world model drives **goal-conditioned neural policies**.
- ▶ Designed **Belief World Tree Search**, an MCTS variant for uncertain maps, with a convergence proof.
- ▶ **30–100× fewer samples**, **20× faster** than the SOTA HRL baseline; **25% fewer steps** than D* on AntMaze.

2025 – 2026
personal

RL Fine-Tuning of Language Models with GRPO

 code

STACK PyTorch · Hugging Face Transformers · TRL · GRPO · GSM8K

- ▶ Implemented **GRPO from scratch** (group-relative advantages, PPO clipping, KL penalty to a reference model) to teach a from-scratch **GPT**, pretrained only on addition, to perform **subtraction**.
- ▶ Applied the same recipe to a pretrained **LLM** (SmolLM2-360M-Instruct) with **TRL** on GSM8K math reasoning; open-sourced as a tutorial series covering **RLHF**, **DPO** and **GRPO**.

2025 – 2026
real hardware

FiTenth Autonomous Race Car: Full ROS 2 Navigation Stack

 code

STACK ROS 2 · Python/C++ · NVIDIA Jetson AGX · LiDAR · RealSense IMU · VESC · SLAM Toolbox · Flask

- ▶ Coded and deployed the race stack on the car's **Jetson AGX**: **SLAM** mapping, **EKF** odometry–IMU fusion, **particle-filter** localisation, and **Pure Pursuit** trajectory tracking.
- ▶ Wrote a LiDAR **time-to-collision emergency-braking** safety node, a gap-following planner, and a Flask **mission-control dashboard** for launch, teleop and odometry tuning.

JMLR 2026
ECAI 2024
published

Differentiable Decision-Tree Policies for Interpretable Deep RL

 code

STACK PyTorch · PPO · SAC · Stable-Baselines3 · CleanRL · Gymnasium

- ▶ Created **DTSemNet**, an exact neural encoding of decision trees trained by **plain gradient descent**; a **drop-in policy network** for RL pipelines.
- ▶ Trained with **PPO** and a modified **SAC** on continuous-control tasks (BipedalWalker, LunarLander), **matching neural policies** while staying human-readable.
- ▶ **Annealed Top- k routing** (MoE-style): beat the best baseline on 80% of regression and all classification tasks.

AAMAS 2027
under review

Robust Hierarchical RL via Distributional Critics (CriOC)

STACK PyTorch · PPO · A2C · QR-DQN · option-critic · MiniGrid · Gymnasium

- ▶ **Distributional option-critic** agent: quantile critics estimate state criticality; a **variational encoder** spreads risk evenly across sub-policies.
- ▶ PPO/A2C variants, 9 benchmarks: **+44% worst-case return** under adversarial perturbation (–12% nominal).

ICLR 2027
under review

Formal Safety Verification of RL Controllers

STACK PyTorch · Gurobi LP/MILP · abstract interpretation · reachability analysis

- ▶ Learn-then-verify pipeline that **formally certifies reach-avoid and safety** of RL policies via LP/MILP.
- ▶ **11–98× smaller certificates** than axis-aligned baselines, incl. a quadrotor controller.

IEEE ICWS 2021
ICSOC 2020/21
published

Deep RL for Resource Allocation in Edge Computing

STACK Python · deep RL · combinatorial optimisation

- ▶ **Deep RL agent for user allocation** and service placement under QoS and workload fluctuations.

Technical Leadership & Mentoring

Ongoing

SafeDAG: Precedence-Constrained LLM Steering for Safer Planning in Embodied Agents

Final-year project supervisor, NTU (ongoing)

- ▶ Steering **LLM planners** with precedence (DAG) constraints so embodied agents produce safe, correctly ordered plans.

2023 – Present

Student Projects & Teams

NTU final-year projects, URECA undergraduate research, FiTenth

- ▶ Supervised **LLM-generated interpretable RL policies**, **ARC with LLMs**, **Decision Transformers**, mechanistic interpretability and CNN robustness.
- ▶ Technical lead of the **FiTenth autonomous racing team**: built the race stack (above) and guided students on control and navigation; VP of CCDS student club (Deep Learning Bootcamp, AI Meetup).

Awards & Publications

Awards

DesCartes Ph.D. Scholarship, CNRS@CREATE (2023) · M.Tech. Scholarship, ISI Kolkata (2019).

Publications

5 first-author peer-reviewed papers: **JMLR** (2026), **ECML-PKDD** (2026), **ECAI** (2024), **SEGAN** (Elsevier, 2024), **IEEE ICWS** (2021); plus 2 at ICSOC (2020, 2021) and 3 under review (ICLR, AAMAS, AAAI).