MPC-SHIELDED REINFORCEMENT LEARNING FOR SAFE AUTONOMOUS EXPLORATION IN SEARCH AND RESCUE

Author: Jokūbas Dimša (j.dimsa@student.tudelft.nl) | Supervised by A. Jamshidnejad & S. Schoonebeek

1. Background and Motivation

- Search and rescue requires fast exploration in dangerous and unknown sites as its too risky for humans, and every minute costs lives.
- Reinforcement Learning explores efficiently and decides in real time, but offers no safety guarantee
- Model Predictive Control is safe by construction, but conservative and slow.
- **The gap:** no method combines RL's fast exploration with MPC's safety under partial observability, where the robot must act on a decaying belief and not on the true state.
- **The approach:** shield the learned policy with an MPC backup, run RL only when it's provably safe under belief.

2. Research Questions

Under what conditions does MPC-shielded RL outperform pure RL and pure MPC on a partially observable SaR task?

- **Learned exploration:** Is RL alone enough to explore a disaster site safely and completely?
- **Safe shielding:** Can we add a safety shield without compromising the RL agent's efficiency?
- **Deployment:** Across safety, effectiveness, and cost, when does each approach win?

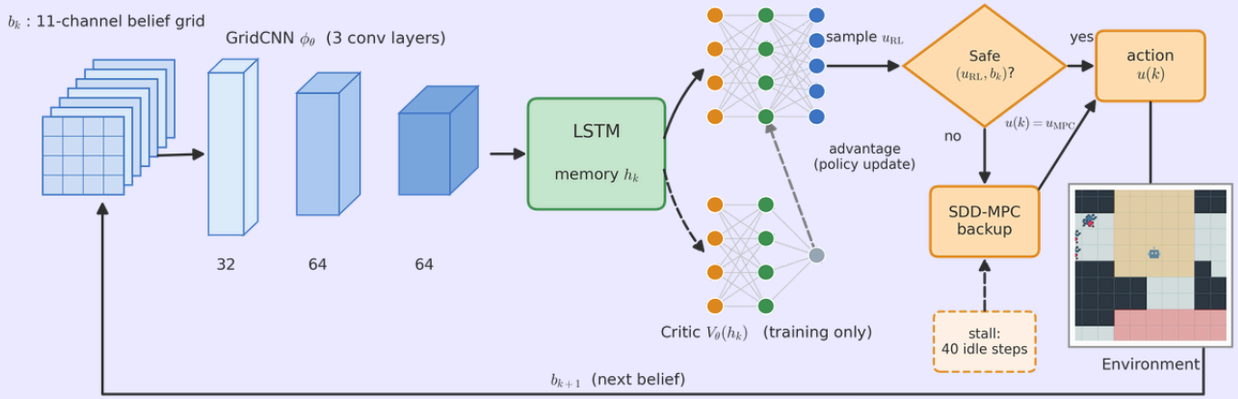
3. Setup & Formalization

- **Agents:** Pure RL, MPC, SDD-MPC, Shield and Shield+C all on 250 shared maps
- **Hazards kept out of the reward** as encoding them made the agent freeze, so safety delegated to the shield
- **Open maps, hazards in rooms** to keep the tube MPC feasible and so the comparison is fair
- **Agent only act on belief, never ground truth**, so it can't use info it doesn't have

$$r_k = w_v\,\Delta V_k + w_{\text{succ}}\,\mathbb{1}^{\text{succ}} + w_{\text{cov}}\,\text{cov}(b_k) - w_s - w_c\,\mathbb{1}^{\text{coll}} + \gamma\Phi(b_{k+1}) - \Phi(b_k)$$

- **Rewards:** finding victims (ΔV), map coverage (cov), completing the task ($\mathbb{1}^{\text{succ}}$)
- **Penalizes:** each step and collisions ($\mathbb{1}^{\text{coll}}$)
- **Shaping term (γΦ(b_{k+1}) – Φ(b_k))** densifies the exploration signal without changing the optimal policy
- **r** is the reward that an agent gains at each step.
- **ΔV_k** if for new victims found.
- **cov(b_k)** is the coverage fraction
- $\mathbb{1}^{\text{succ}}$ when all victims found.
- $\mathbb{1}^{\text{coll}}$ when there is a collision (either wall or victim).
- **Φ** is the coverage shaping.
- **γ** is the discount factor.
- **w•** weights

4. Methodology

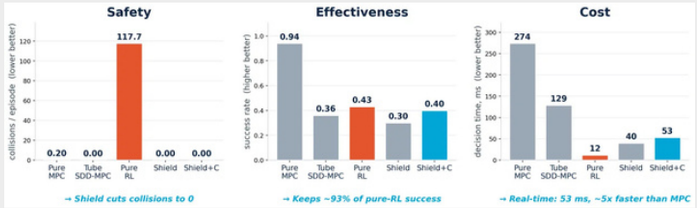


The agent's 11 channel belief is compressed by a CNN into a 256-feature vector, which is carried through an LSTM for memory under partial observability and then read by an actor and critic (critic in training only). The actor outputs a softmax over the 5 actions (up, down, left, right, wait) and the shield only executes it only if its safe, otherwise the SDD-MPC backup takes over.

LSTM quadruples success (0.12 → 0.50):

Policy	Succ	Cov	Vic
Frame-stack PPO	.65	.66	.12
LSTM (recurrent)	.75	.79	.50

5. Results



Agent	Succ	Cov	Vic	Coll	Dmg	ms
MPC	**.94**	**.79**	**.95**	0.2	**6.5**	274
Tube SDD-MPC	.36	.57	.63	**0**	11.9	129
Pure RL	.43	.69	.77	117.7	175.8	**12**
Shield	.30	.64	.71	**0**	12.0	40
Shield+C	.40	.64	.72	**0**	12.1	53

- The shield eliminates RL's collisions while keeping most of its success and speed.
- Among the collision-free agents, Shield+C explores more than SDD-MPC and decides 2.4x faster.
- MPC is most effective with 94% success rate, but its 5x slower and cannot scale well.
- Wilcoxon (250 paired seeds): Shield+C ties pure RL and SDD-MPC on success. Its coverage, victim and speed gains over SDD-MPC are significant ($p < 0.01$).

6. Conclusion

- **Learned exploration:** RL alone isn't enough, it hits a ceiling near 0.6 coverage and collides heavily (~117 times an episode), memory (LSTM) lifts this up to 0.75 but still unsafe.
- **Safe shielding:** the shield that I present, adds safety without losing efficiency. Collision reduce down to 0 and success similar to RL at a real-time speed (53ms).
- **Deployment:** no single winner exists, but the shield wins when safety is critical and decisions must stay fast. The exhaustive MPC wins only when raw effectiveness outweighs speed and scalability.

7. Future Work

- **Size-independent policy:** replace the whole-map CNN with an egocentric crop or global pooling to generalize across grid sizes.
- **Hierarchical planning:** add high-level planning to push past the ~0.7 coverage ceiling.
- **Towards real robots:** add richer disturbances and validate the slip model on hardware.
- **Multi-agent rescue:** coordinate multiple robots and reward rescuing victims, not just seeing them.