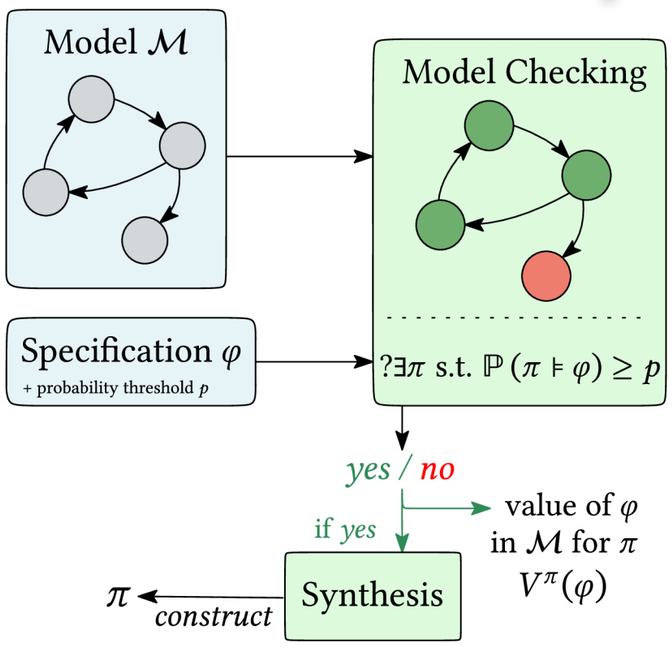
Composing Reinforcement Learning Policies, with Formal Guarantees

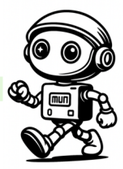
Florent Delgrange, *Guy Avni, Christian Schilling, Anna Lukina, Ann Nowé, Guillermo A. Pérez*

Reactive Synthesis

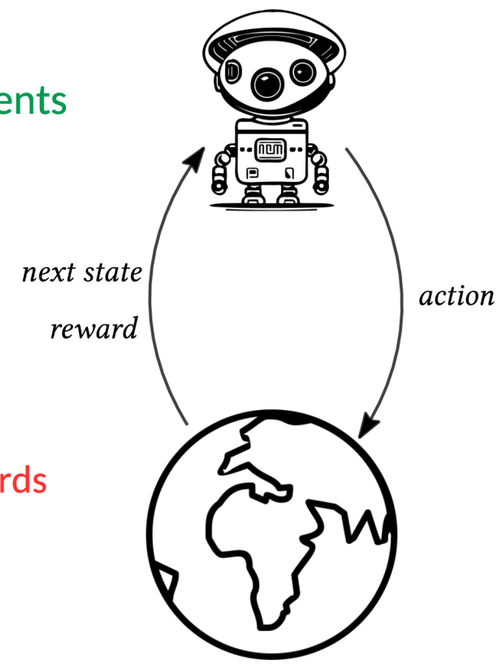


- [-] Scalability issues
- [-] Requires access to a **model** of the environment
- [+] Specification Formalisms
  (e.g., LTL, PCTL, μ-calculus, etc)
- **[+] Formal Guarantees**

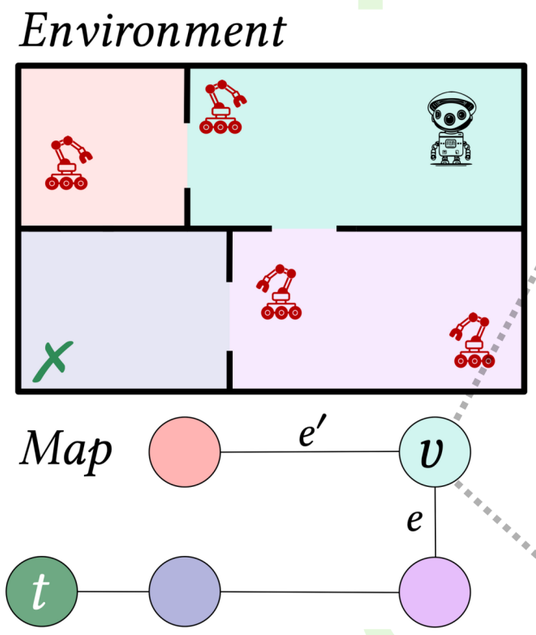
Reinforcement Learning

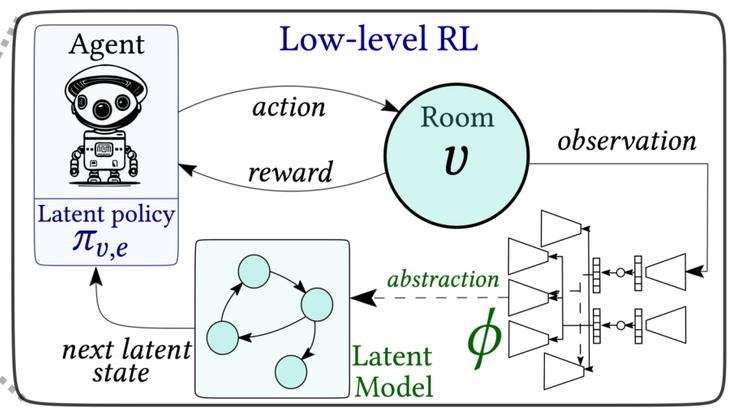


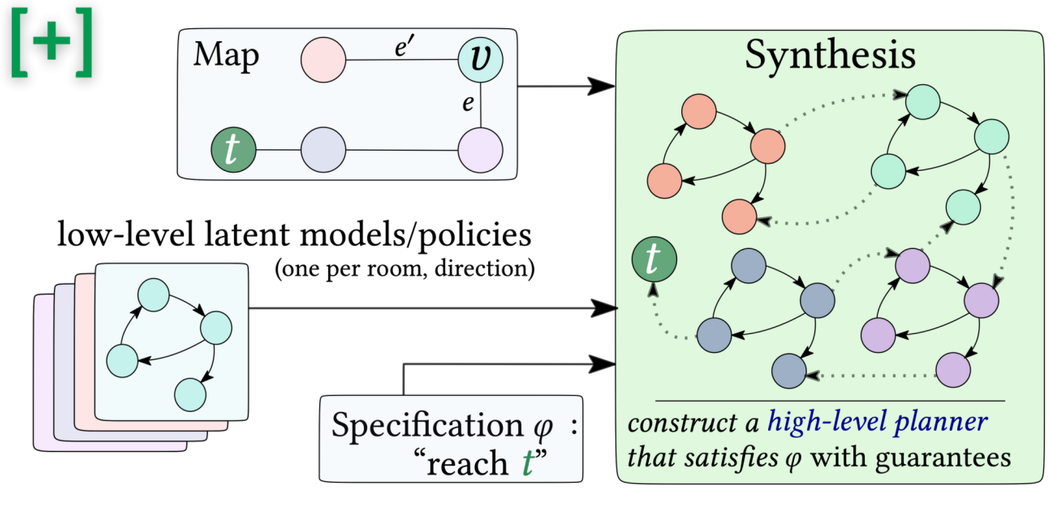
- [+] Scales to complex environments
- [+] Interacting with the environment is sufficient
- [-] Requires designing a reward function
  → reward engineering, sparse rewards
- **[-] No Guarantee**



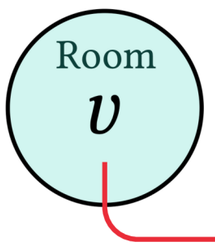
Our goal: get the best [+] of both worlds

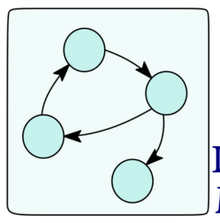






Formal Guarantees





- Learn $\overline{\mathcal{M}}$ via a **loss**:

$$L_{\mathbf{P}}^{v,e} = \mathbb{E}_{s,a\sim\pi_{v,e}} \mathrm{TV}\left(\phi_{\sharp}\mathbf{P}(\cdot \mid s,a), \overline{\mathbf{P}}(\cdot \mid \phi(s),a)\right)$$

- Room-wise **Abstraction Quality Guarantees**:

$$\left|V_{v}^{\pi_{v,e}}(\varphi) - V_{\overline{\mathcal{M}}}^{\pi_{v,e}}(\varphi)\right| \leq \mathrm{AEL}(\pi_{v,e}) \cdot \frac{L_{\mathbf{P}}^{v,e}}{1-\gamma}$$

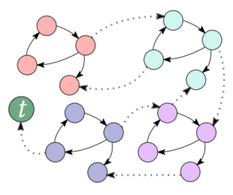
PAC Learnable

Average Episode Length

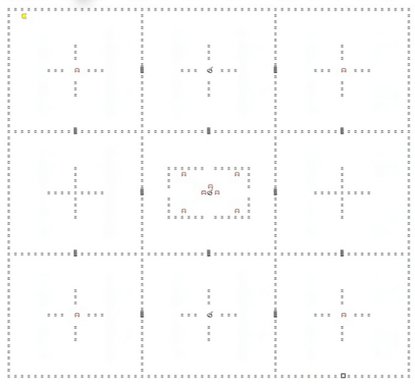
- Memory is **necessary** for the HL planner τ
- Optimal τ obtained by solving an MDP
- **Lifting the guarantees** to the high-level:

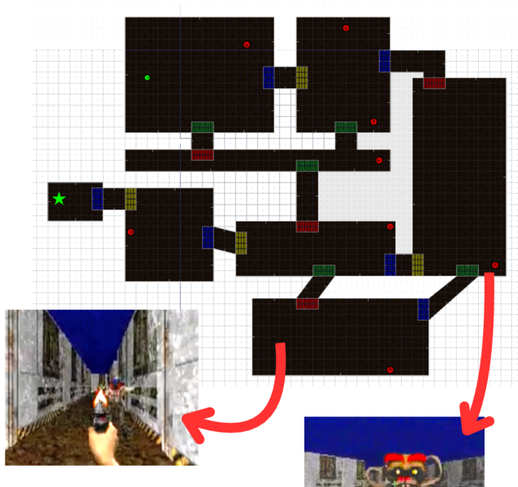
$$\left|V_{\text{true}}^{\tau}(\varphi) - V_{\mathcal{C}}^{\tau}(\varphi)\right| \leq \mathrm{AEL}(\tau) \cdot \frac{L_I + K \cdot \mathbb{E}_{(v,e)\sim\tau}\left[L_{\mathbf{P}}^{v,e}\right]}{1-\gamma}$$

Composed MDP $\mathcal{C}$



Experiments



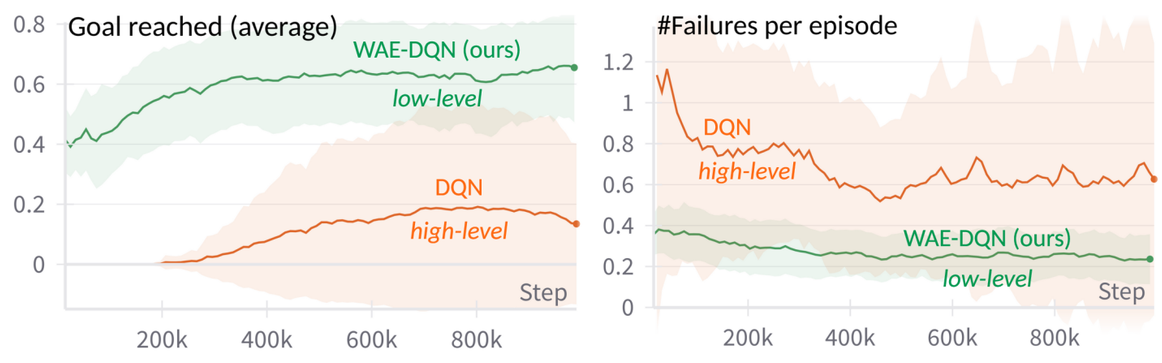


N: #rooms, *LP*: life points, *A*: #adversaries

	N	LP	A	avg. return ($\gamma = 1$)	latent value	avg. value (original)
Grid World	9	1	11	0.5467 ± 0.1017	0.1378	0.07506 ± 0.01664
	9	3	11	0.7 ± 0.09428	0.4343	0.01 ± 0.00163
	25	3	23	0.4933 ± 0.09832	0.1763	0.007833 ± 0.002131
	25	5	23	0.5667 ± 0.07817	0.346	0.00832 ± 0.00288
	49	7	47	0.02667 ± 0.01491	0.004229	5.565e-6 ± 7e-6
ViZDoom	8	/	8	0.89333 ± 0.059628	0.24171	0.23405 ± 0.014781
	8	/	14	0.78 ± 0.064979	0.16459	0.16733 ± 0.023117
	8	/	20	0.39333 ± 0.11643	0.086714	0.06898 ± 0.017788

e	Grid World	ViZDoom
→	0.50412	0.32011
←	0.77787	0.44883
↑	0.49631	0.37931
↓	0.48058	0.48108

Table 2: PAC bounds.



check out the paper!

Paper

Blogpost

in/florent-delgrange
@florentdelgrange.bsky.social
@f_delgrange
https://delgrange.me/

VUB ARTIFICIAL INTELLIGENCE RESEARCH GROUP
TU Delft Delft University of Technology
Universiteit Antwerpen
AALBORG UNIVERSITY