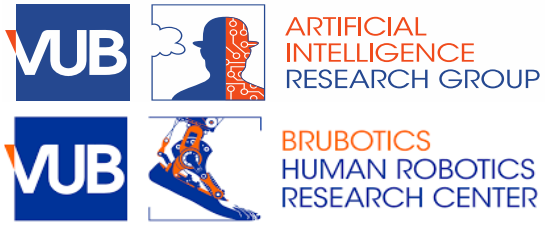



# $LTL_f$-based Reward Shaping for Reinforcement Learning

Mahmoud Elbarbari, Kyriakos Efthymiadis, Bram Vanderborght, Ann Nowe

WORK-IN-PROGRESS

**This work is part of the Flanders AI research program**
**www.airesearchflanders.be**

# MOTIVATION

- Real-world tasks in partially unkown environments e.g., delivering parts in a machine tending task.
- RL provides an appealing solution (but slow).
- In the literature, there are different methods to speed up RL.

# MOTIVATION

Potential-based
Reward Shaping

# MOTIVATION

(non-expert)
User

Potential-based
Reward Shaping

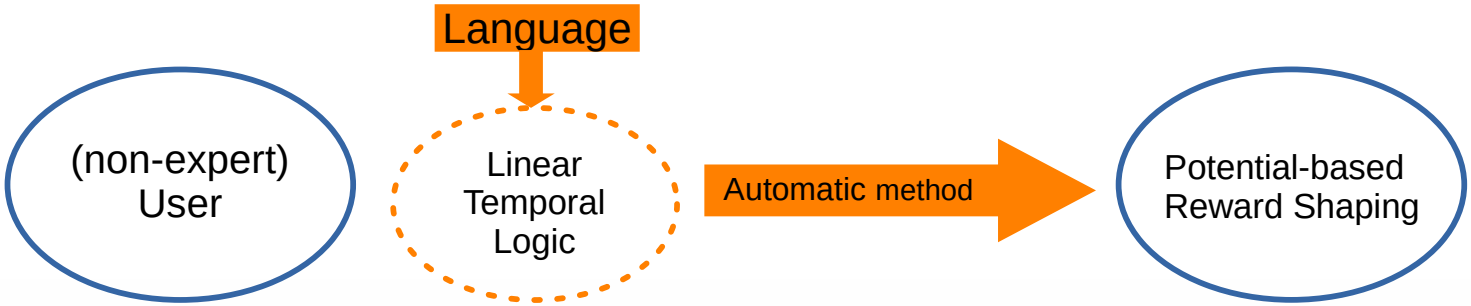
# MOTIVATION


Language
(non-expert)
User
Linear
Temporal
Logic
Potential-based
Reward Shaping

# MOTIVATION


- Similarity to natural language.
- Richness and compactness.
- Flexibility.

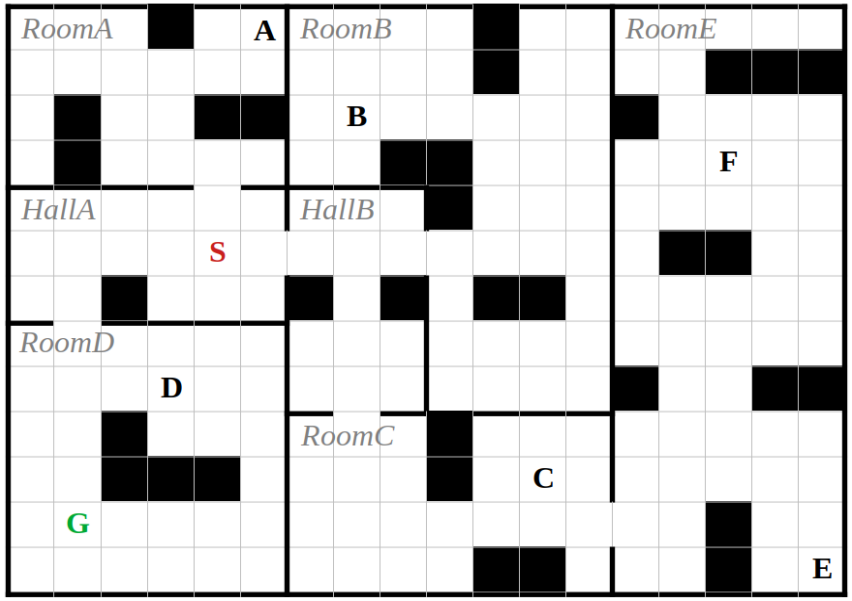
# MOTIVATION




- Similarity to natural language.
- Richness and compactness.
- Flexibility.

# MOTIVATION

## Classic flag collection domain



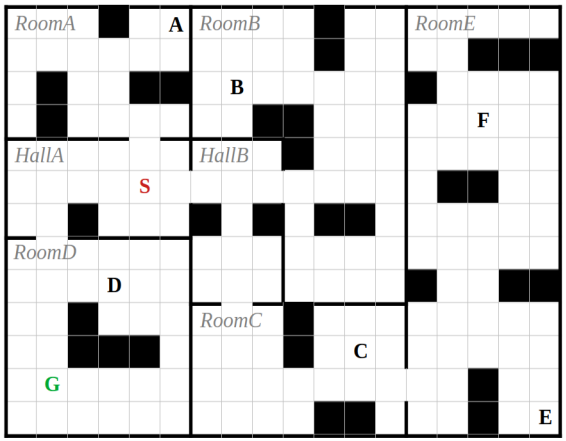

*Flag collection domain (Grzes and Kudenko, 2008).*

# MOTIVATION

## Classic flag collection domain

- Using $LTL_f$ to guide the agent to an optimal solution (A, B, C, E, F then D).

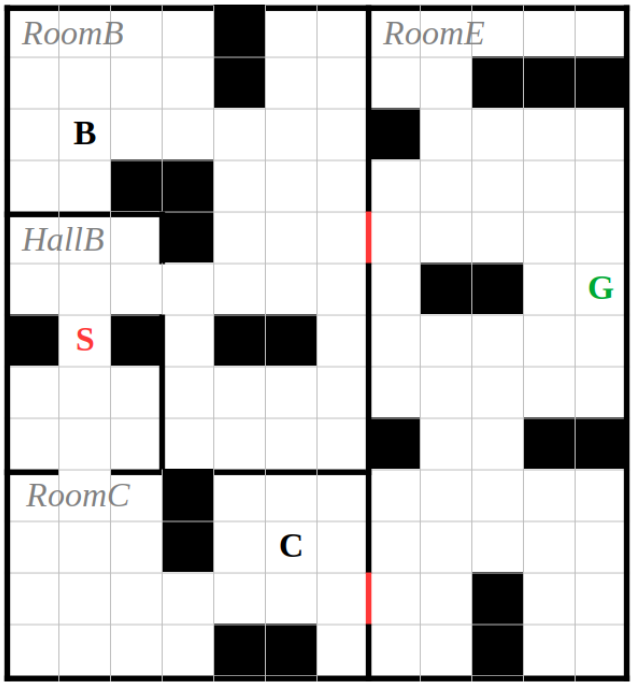
$$F\left(have\_flagA \wedge X\, F\left(have\_flagB \ldots \wedge X\, F\, at\_goal\right)\right)$$




*Flag collection domain (Grzes and Kudenko, 2008).*

# EXPERIMENTS

## Flag collection domain with partial knowledge

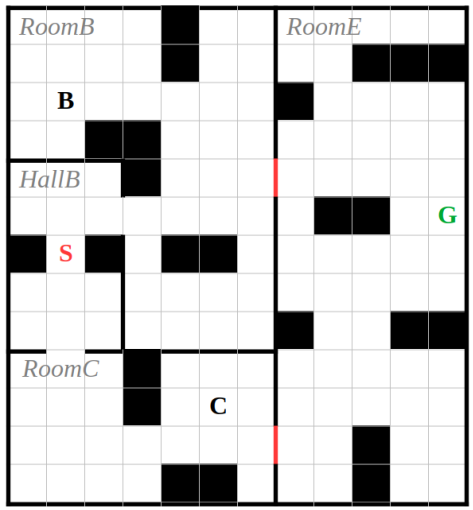


# EXPERIMENTS

## Flag collection domain with partial knowledge

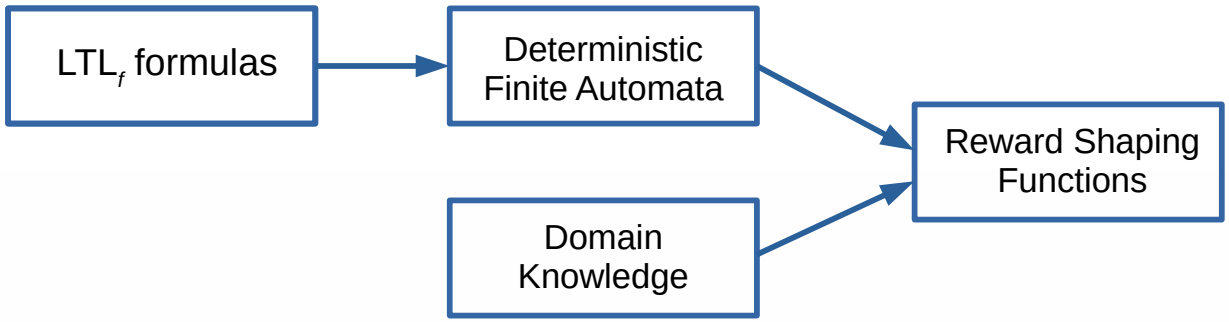
- Guide the agent to collect flag B and C (without specifing an order).

$$F((have\_flagB \wedge have\_flagC) \wedge XFat\_goal))$$



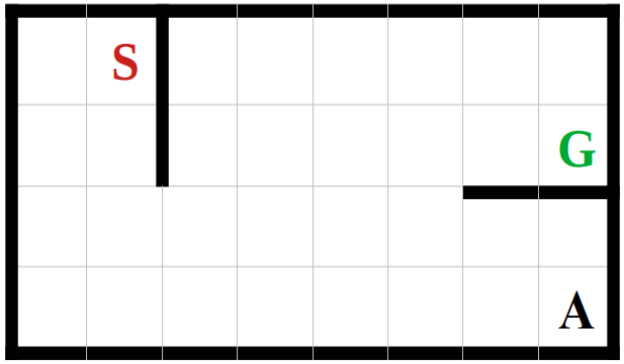
# APPROACH

## Overview



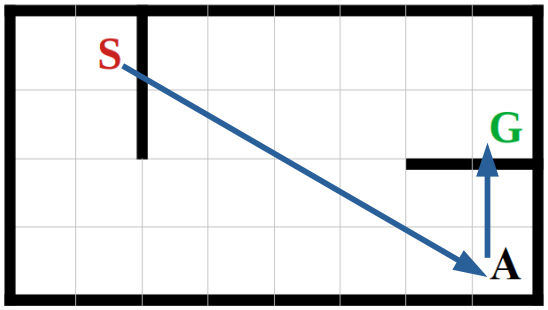
# APPROACH

## Running example



# APPROACH

## LTL$_f$ formula

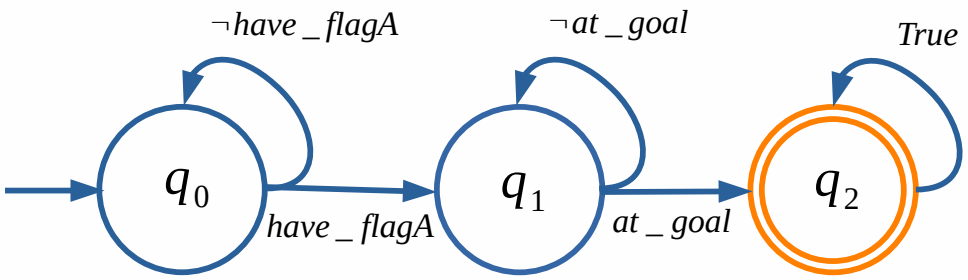



$$F\left(have\_flagA \wedge X\, F\, at\_goal\right)$$

# APPROACH

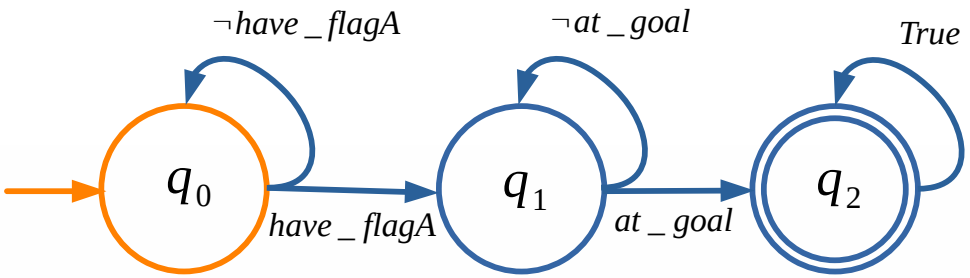
## From $LTL_f$ to DFA

$$F\left(have\_flagA \wedge X\, F\, at\_goal\right)$$



# APPROACH

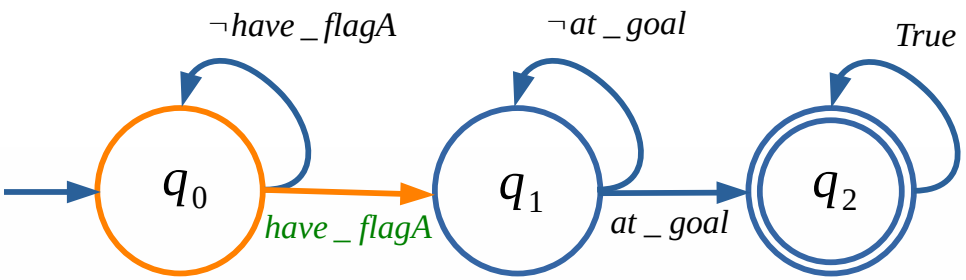
## The guiding formula



# APPROACH

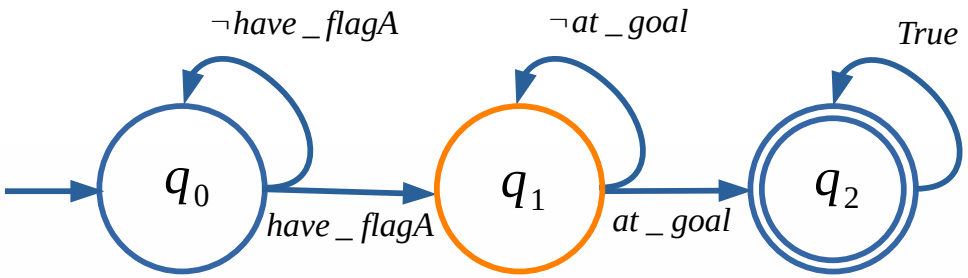
## The guiding formula

$$\phi^{t}_{guiding} = have\_flagA$$



# APPROACH

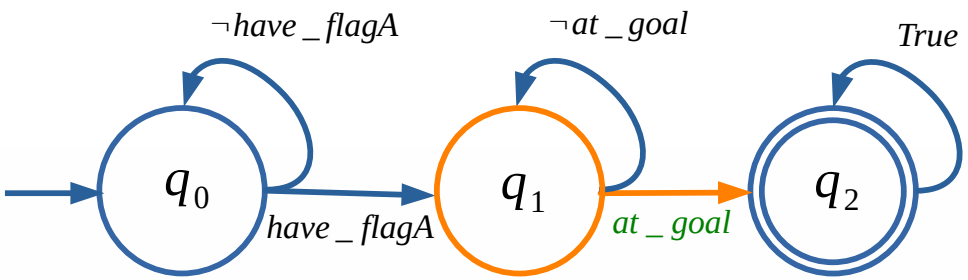
## The guiding formula



# APPROACH

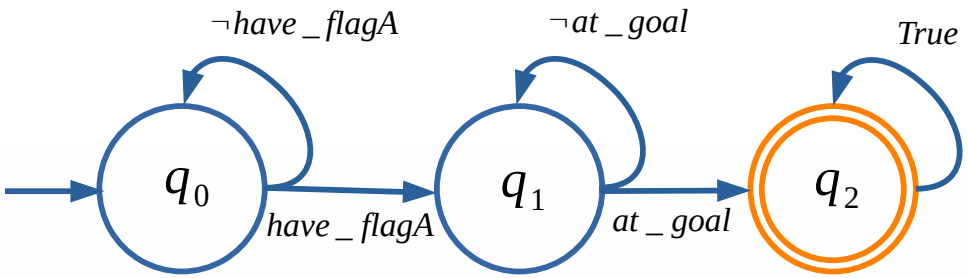
## The guiding formula

$$\phi^{t}_{guiding} = at\_goal$$



# APPROACH

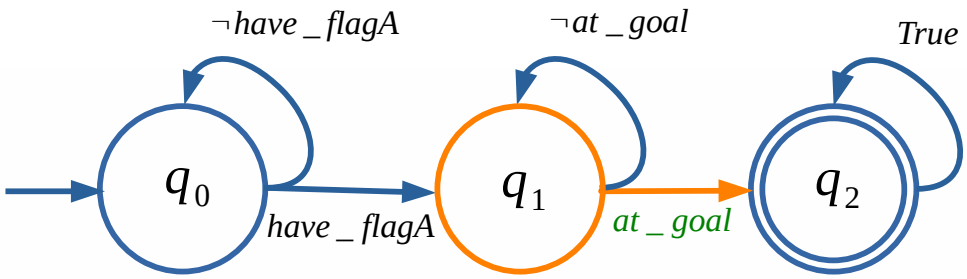
## The guiding formula

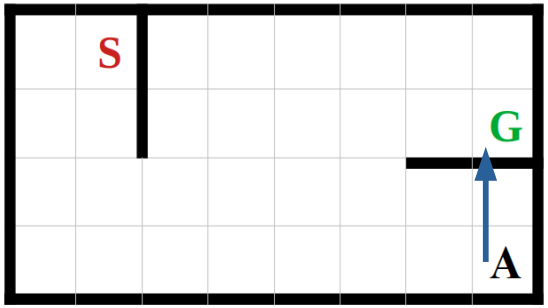


# APPROACH

## Domain knowledge

$$h(s_t, \phi^t_{guiding})$$




$$h(s_t, at\_goal)$$

# APPROACH

## The potential function

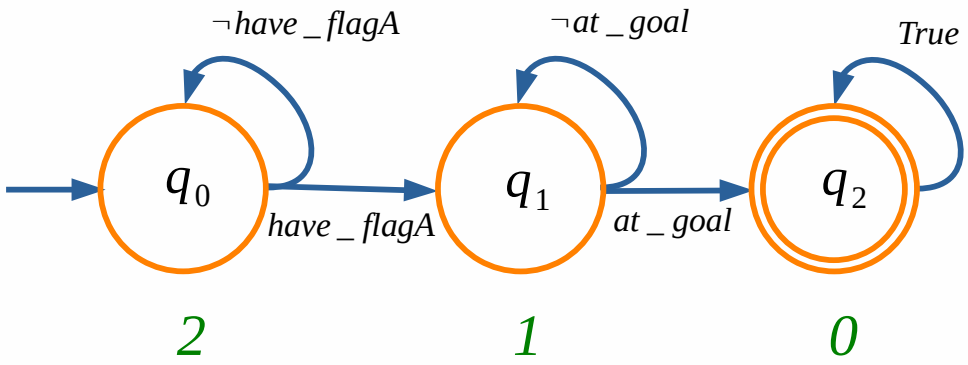
$$\Phi_t(s_t) = -\omega \times [Distance(q_t) + \frac{1}{n} \times h(s_t, \phi^t_{guiding})]$$

$$n = max(h(S, \phi^t_{guiding}))$$

APPROACH

## The potential function

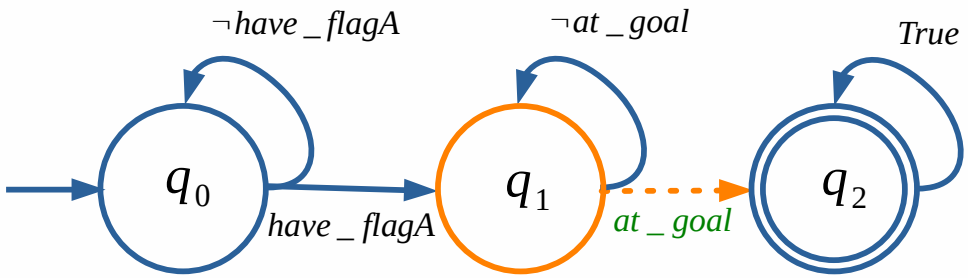
$$\Phi_t(s_t) = -\omega \times [\, Distance(q_t) + \frac{1}{n} \times h(s_t, \phi^t_{guiding}) \,]$$



APPROACH

## The potential function

$$\Phi_t(s_t) = -\omega \times [\, Distance(q_t) + \frac{1}{n} \times h(s_t, \phi^t_{guiding}) \,]$$



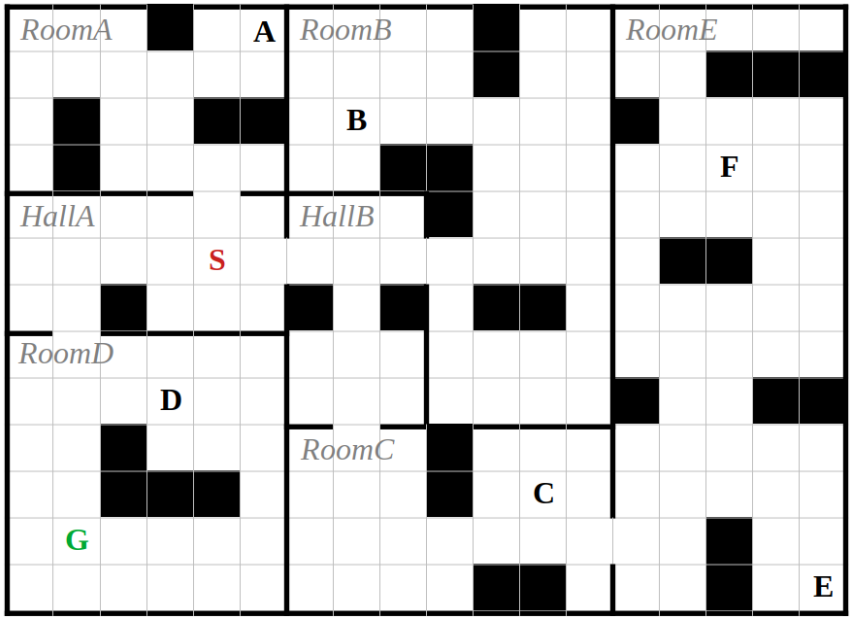
# EXPERIMENTS

## Plan-based reward shaping

- (Grzes and Kudenko 2008) proposes a method to automatically generate PBRS functions from STRIPS plans.

# EXPERIMENTS

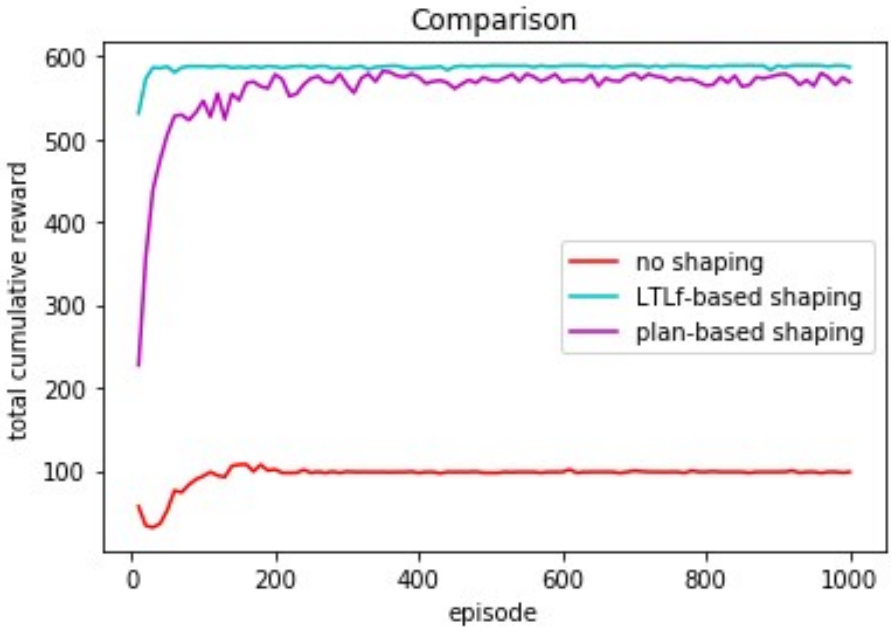
## Classic flag collection domain




- Multiple optimal orderings.
- Plan-based guides to one optimal ordering.
- $LTL_f$-based provides flexibility to guide to any optimal ordering.

# EXPERIMENTS

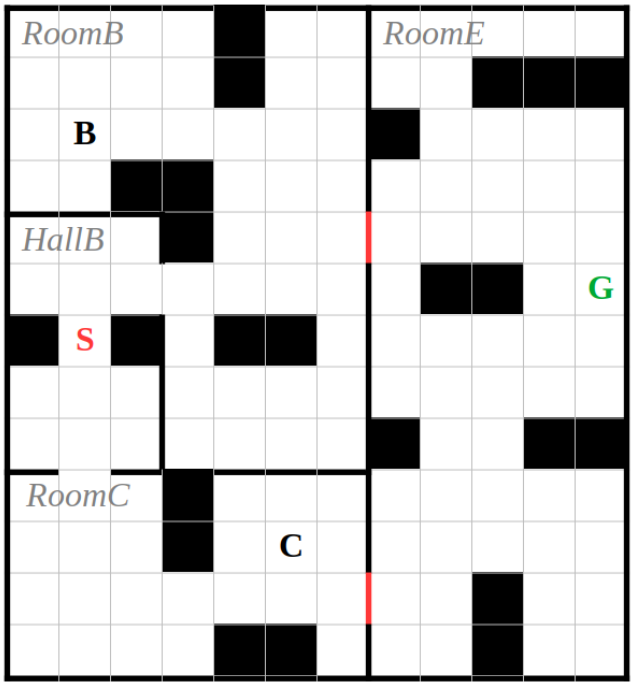
## Classic flag collection domain




- Guiding to same optimal ordering.
- $LTL_f$ achieves better performance.
- Using $LTL_f$ for guiding to other optimal ordering reports similar results.
- No shaping converges to suboptimal solution.

# EXPERIMENTS

## Flag collection domain with partial knowledge



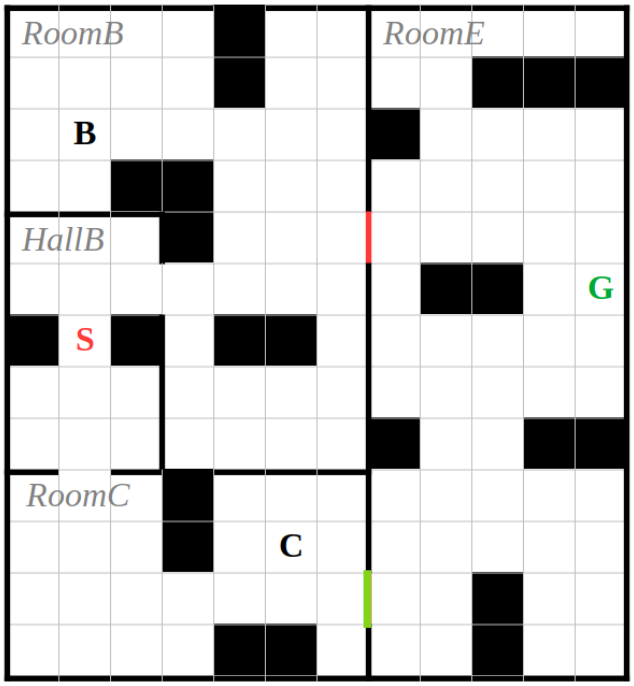

- Flexible advice

$$F((have\_flagB \wedge have\_flagC) \wedge XFat\_goal))$$

- Plan-based method can not provide such flexibility.

# EXPERIMENTS

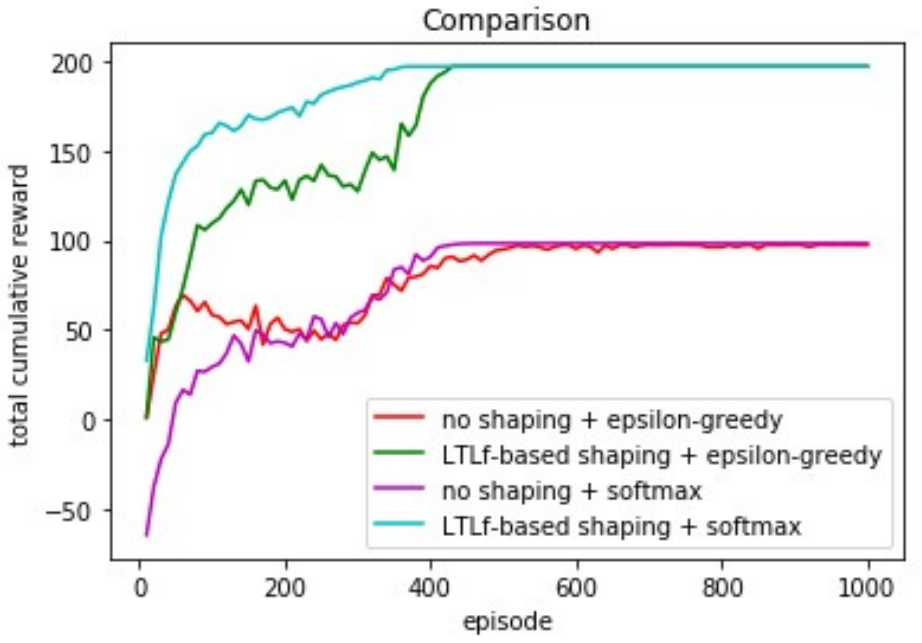
## Flag collection domain with partial knowledge




- The door (RoomC-RoomE) is the open door.
- Optimal ordering: flag B then C.

# EXPERIMENTS

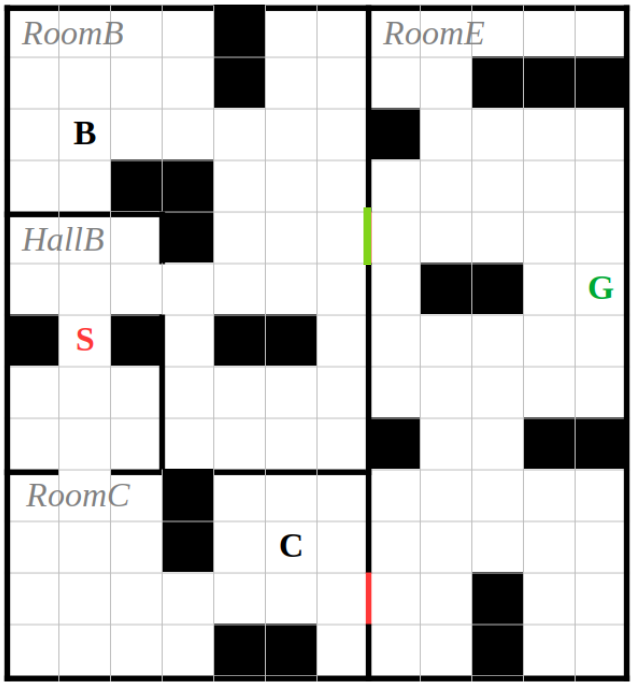
## Flag collection domain with uncertainty




- We compare $LTL_f$ and no shaping.
- We use epsilon-greedy and softmax.
- Softmax achieves better initial performance.

# EXPERIMENTS

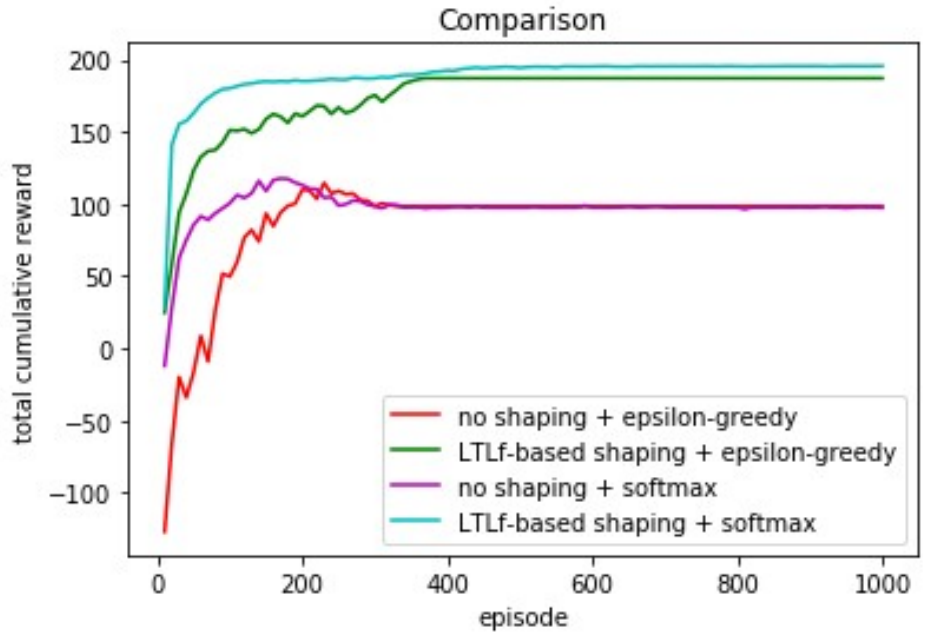
## Flag collection domain with uncertainty




- The door (RoomB-RoomE) is the open door.
- Optimal ordering: flag C then B.

# EXPERIMENTS

## Flag collection domain with uncertainty




- Epsilon-greedy fails to learn the optimal ordering.

# CONCLUSION

- We introduced a novel methodology to generate PBRS functions using user-provided $LTL_f$ formulas.

- We demonstrated
  - better performance.
  - aspects of flexibility.

# FUTURE WORK

- Further demonstrate $LTL_f$ advantages.
- Investigating the use of $LTL_f$ in multi-agent systems.
- Extending our experiments to real robotic systems.

# THANK YOU

**CONTACT**
*mahmoud.elbarbari@vub.be*

VRIJE
UNIVERSITEIT
BRUSSEL