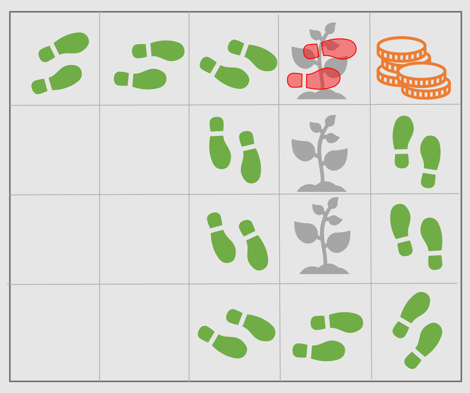
Planning for Potential: Efficient Safe Reinforcement Learning

Floris den Hengst[1,2]
Floris.den.Hengst@ing.com
Vincent François-Lavet[1]
Mark Hoogendoorn[1]
Frank van Harmelen[1]

Introduction

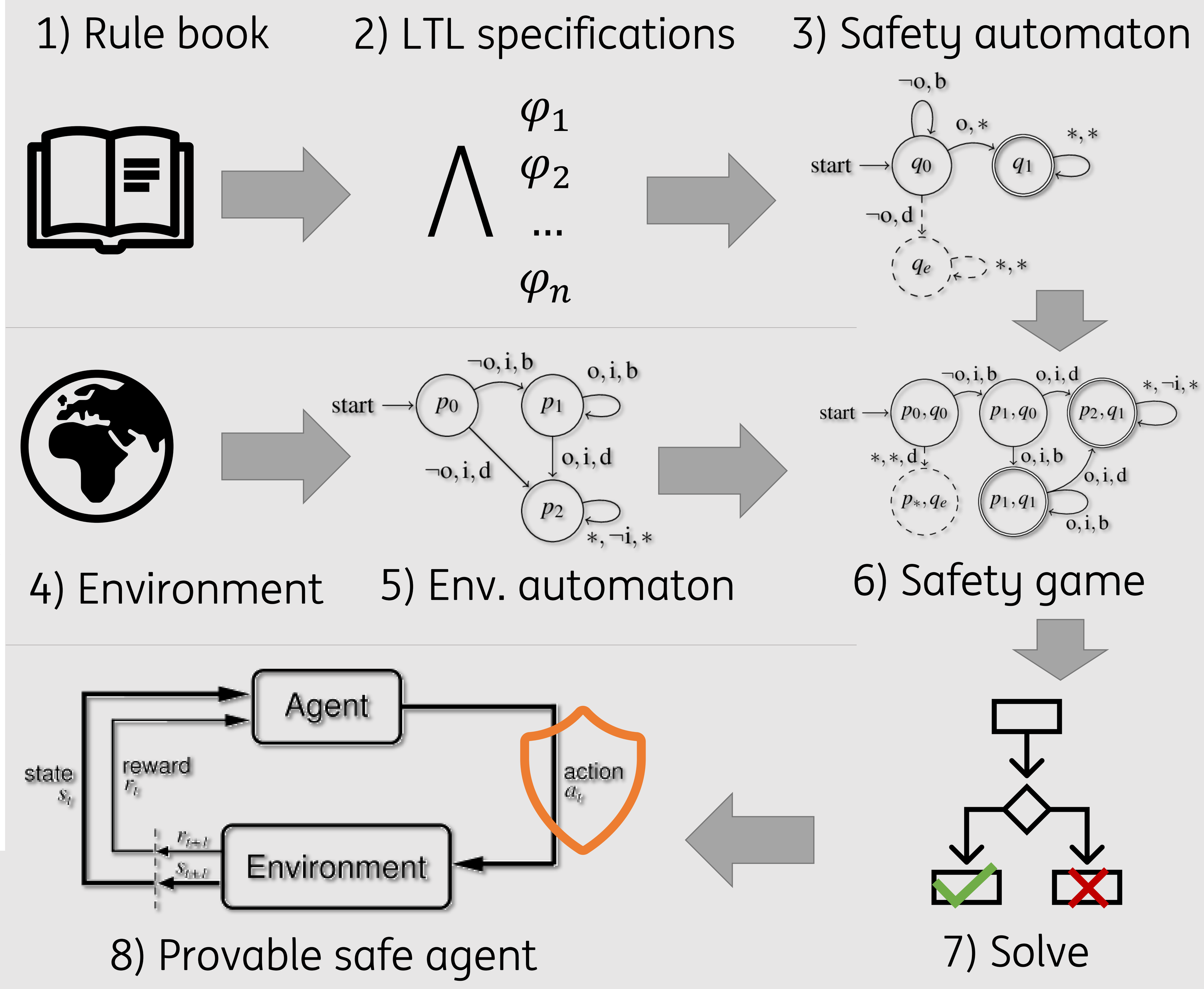
Safety constraints that should never be violated in e.g. healthcare and finance

Conditions that have to be met at all times
- Symbolic
- Temporal
- High-level



Learning while safe: easier or harder?
- being safe ⇒ high reward?
- being safe ⇒ sparse reward?

RL with Temporal Constraints[0]

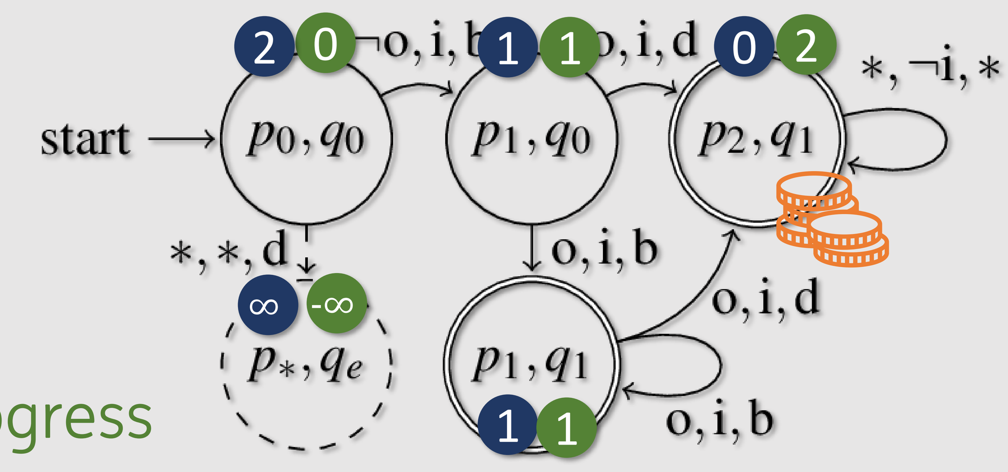


Planning for Potential

Distance metric
$\Delta(p) \coloneqq$ # steps in automaton towards high-level goal

Actions that reduce this distance are valued higher under π^*

Algorithm sketch:
1. For every automaton state p_x:
2. Compute distance $\Delta(p_x)$
3. Compute progress $\Delta(p_0) - \Delta(p_x)$
4. Assign potentials $\phi(p_x) \coloneqq c\,(\Delta(p_0) - \Delta(p_x))$
5. For every time step t:
6. Generate (s, p, a, r, s', p')
7. Shape reward $r' \coloneqq r + \gamma\phi(p') - \phi(p)$[1]
8. Update π with (s, a, r', s)



Discussion

Relation between distance and reward for safe RL

Scale safe RL by learning and reasoning over constraints

Inform learner of progress with potential-based shaping

New Questions:
- Some constraints have a large impact. Why? Identifiable a priori?
- What if the constraints change?
- Can we learn the environment model/automaton?[3]
- Beyond safety: prior knowledge as constraints?

Experiments

Algorithms

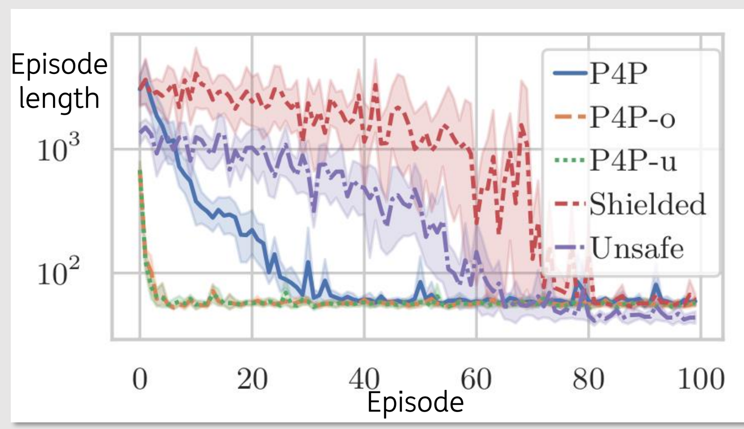
P4P-a: tune c on the fly
P4P-o: overestimate c
P4P-u: underestimate c

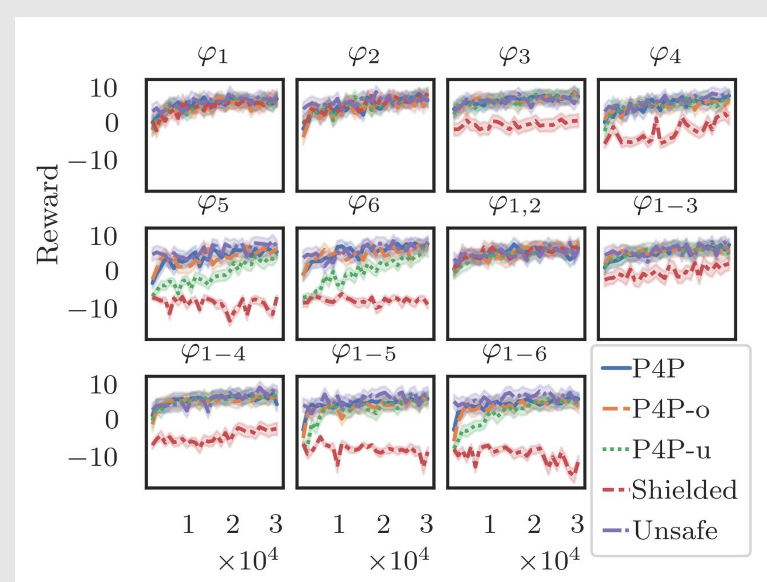
Shielded: safe RL baseline
Unsafe: vanilla RL baseline

Environments

Grid world
from literature
toy example
tabular Q-learning

Chatbot[3]
real-world constraints
learned simulator
DQN





P4P significantly outperforms safe baselines

Results comparable to unsafe baselines <- P4P nullifies costs for being safe!

Scalable: performance stable as problem is more constrained

Robust with respect to hyperparameter c
- Can be tuned automatically
- Set up front using domain knowledge
- OK if 'poorly' chosen

[0] Alshiekh, Mohammed, et al. "Safe reinforcement learning via shielding." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. No. 1. 2018.
[1] Ng, Andrew Y., Daishi Harada, and Stuart Russell. "Policy invariance under reward transformations: Theory and application to reward shaping." Icml. Vol. 99. 1999.
[2] Den Hengst, Floris, et al. "Reinforcement learning for personalized dialogue management." IEEE/WIC/ACM International Conference on Web Intelligence. 2019.
[3] Verginis, Christos, et al. "Joint Learning of Reward Machines and Policies in Environments with Partially Known Semantics." arXiv preprint arXiv:2204.11833 (2022).