

# Standing Still Is Not An Option

New Baselines for Impact Regularization using  
Attainable Utility Preservation\*

Fabian Kovac

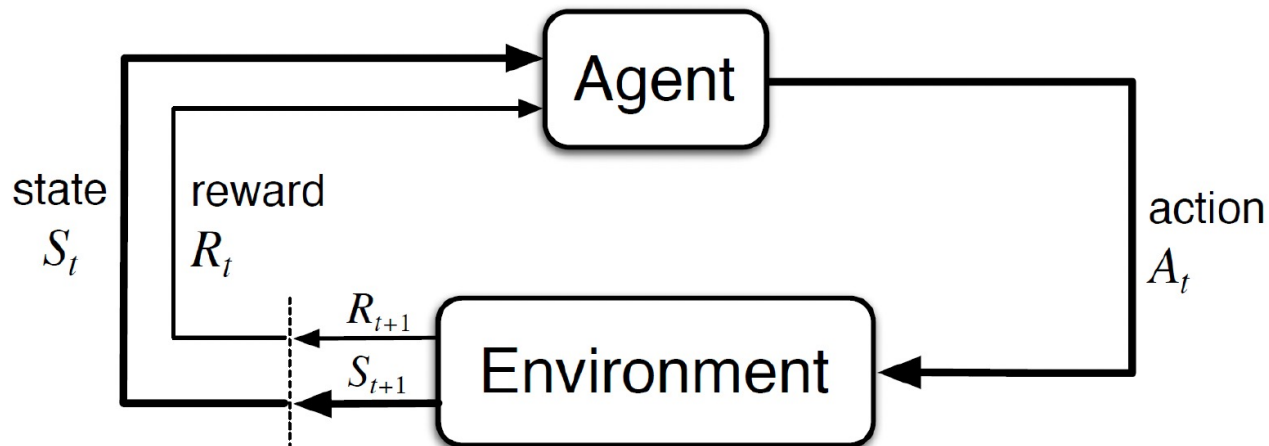
< *fabian.kovac@fhstp.ac.at* >

*St. Pölten University of Applied Sciences, Austria*

\* Work under review at International IFIP Cross Domain (CD) Conference for Machine Learning & Knowledge Extraction (MAKE) CD-MAKE 2023

# Introduction

- Unlike (un)supervised learning, Reinforcement Learning does not focus on data itself
  - *Agent* and *Environment* stand in a cyclic relationship with each other
  - *Agent* observes states and rewards from the *Environment* and takes actions

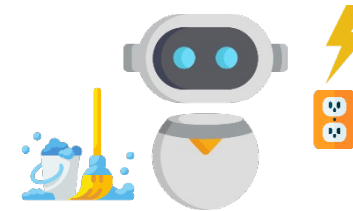
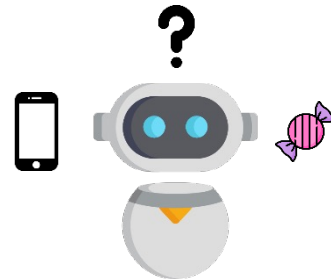
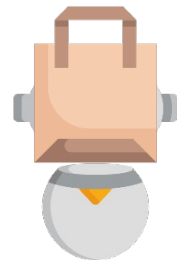
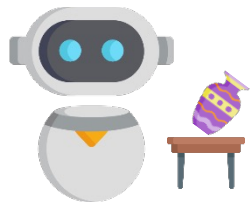


*Agent-environment interaction in Reinforcement Learning [SB18]*

# Introduction

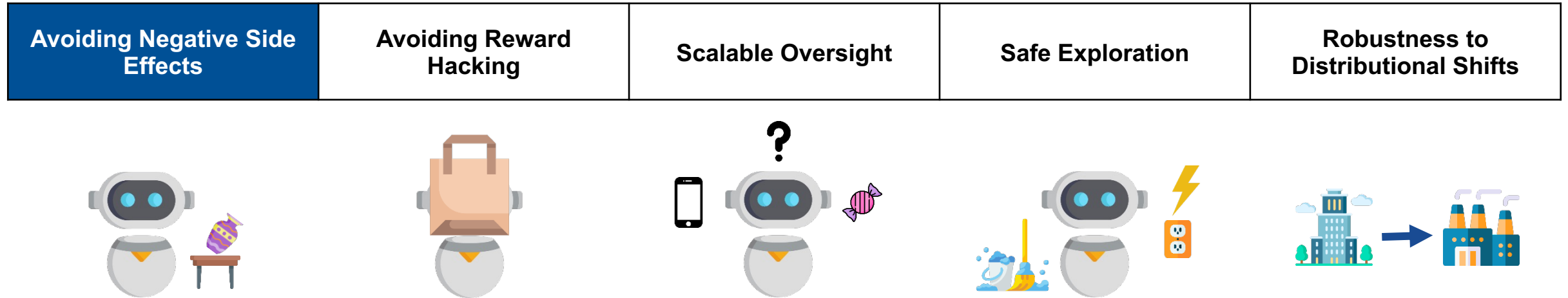
- AI Alignment / AI control problem: *[Gab20; Chr21]*
  - Aspects, on how AI systems should be built, that their preferences align with human values
- Concrete Problems in AI safety regarding Reinforcement Learning: *[Amo+16]*

Avoiding Negative Side Effects	Avoiding Reward Hacking	Scalable Oversight	Safe Exploration	Robustness to Distributional Shifts
--------------------------------	-------------------------	--------------------	------------------	-------------------------------------



# Introduction

- AI Alignment / AI control problem: *[Gab20; Chr21]*
  - Aspects, on how AI systems should be built, that their preferences align with human values
- Concrete Problems in AI safety regarding Reinforcement Learning: *[Amo+16]*



# Related Work

- **Constrained Markov Decision Processes** *[AI99]*
  - Whitelisted constraint themes to avoid side effects *[ZDS18]*
  - Reasonable feasible set for robust rewards *[RB10]*
- **Safe RL** *[PS14; GF15; Ber+17; Cho+18]*
  - Avoiding irreversable mistakes during training
- **Attainable Utility Preservation (AUP)** *[THT20; TRT20]*

# Scope of work

## Unsolved challenges

Current approaches of AUP assume the existence of a no-op action ( $\emptyset \notin \mathcal{A}$ ), which depending on the environment, cannot be always guaranteed.

## Research question

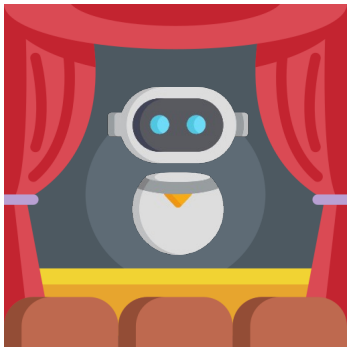
*“ How can Attainable Utility Preservation be extended to single agent environments without no-op actions with a discrete action space? ”*

## Scientific methods

Qualitative (explorative) research methods including literature research and experiments

# Attainable Utility Preservation (AUP)

## Primary objective



primary reward function  $R$

correlates with !?

auxiliary reward functions  $R_i$

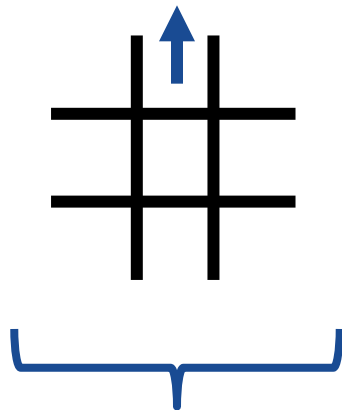
## Auxiliary objectives



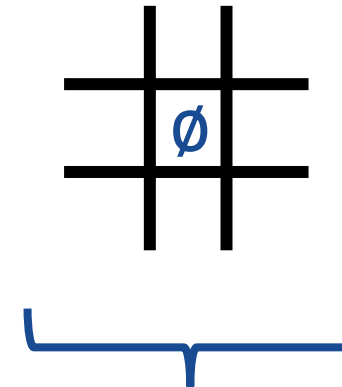
*You, listening to me  
right now ....*

*.... and all the things you  
actually want to do!*

- AUP learns by combining the primary world with auxiliary worlds
  - Penalize primary reward by using auxiliary action-values compared to no-op action (e.g., standing still) to estimate unknown, side-effect free reward



chosen action



no-op action  
(e.g., standing still)



- We consider a finite Markov Decision Process (MDP)  $\langle \mathcal{S}, \mathcal{A}, T, R, \gamma \rangle$  with state space  $\mathcal{S}$ , action space  $\mathcal{A}$ , transition function  $T: \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$ , reward function  $R: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  and a discount factor  $\gamma \in [0, 1)$
- Assumptions:
  - Existence of a no-op action  $\emptyset \in \mathcal{A}$
  - Finite set of random, auxiliary reward functions  $\mathcal{R} \subset \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$ 
    - Each  $R_i \in \mathcal{R}$  has a corresponding  $Q$ -function  $Q_{R_i}$
    - Correct reward function may NOT belong to  $\mathcal{R}$

**AUP penalty:** Let  $s$  be a state and  $a$  be an action

$$PENALTY(s, a) := \sum_{R_i \in \mathcal{R}} |Q_{R_i}(s, a) - Q_{R_i}(s, \emptyset)|$$

**AUP scale:** Normalize penalty to agent's situation

$$SCALE(s) := \sum_{R_i \in \mathcal{R}} Q_{R_i}(s, \emptyset)$$

## AUP reward function: Let $\lambda \geq 0$

Similar to regularization in supervised learning,  $\lambda$  controls the influence of the AUP penalty on the reward function.

$$R_{AUP}(s, a) := R(s, a) - \frac{\lambda}{\mu} \sum_{R_i \in \mathcal{R}} \text{PENALTY}(s, a), \quad \text{where } \mu := \begin{cases} \text{SCALE}(s) & \text{or} \\ |\mathcal{R}| \end{cases}$$

1. Update auxiliary action-value functions
2. Standard  $Q$ -Learning using  $R_{AUP}$  instead of observed reward

---

### Algorithm: AUP update [THT20]

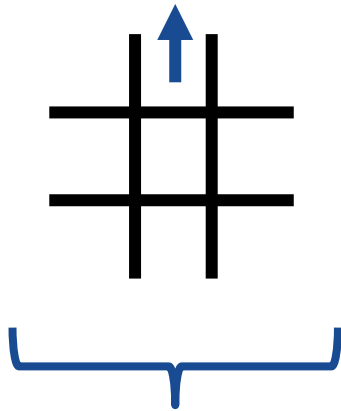
---

```

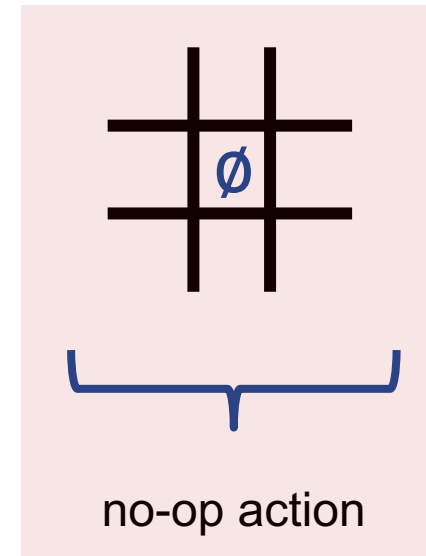
for  $i \in |\mathcal{R}| \cup \{AUP\}$  do
     $Q_{R_i}(s, a) = Q_{R_i}(s, a) + \alpha(R_i(s, a) + \gamma \max_{a'} Q_{R_i}(s', a') - Q_{R_i}(s, a))$ 
     $Q_{AUP}(s, a) = Q_{AUP}(s, a) + \alpha(R_{AUP}(s, a) + \gamma \max_{a'} Q_{AUP}(s', a') - Q_{AUP}(s, a))$ 
    
```

---

- What if "standing still" is not an option?



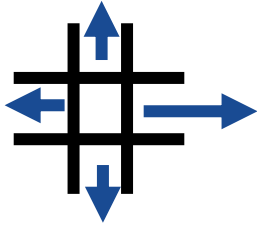
chosen action



no-op action

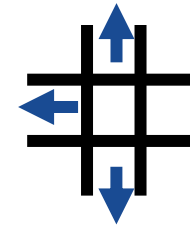
**vAUP (mean):** Penalize compared to mean estimate

$$PENALTY_{mean}(s, a) := \sum_{R_i \in \mathcal{R}} |Q_{R_i}(s, a) - \left( \frac{1}{|\mathcal{A}|} \sum_{a' \in \mathcal{A}} Q_{R_i}(s, a') \right)|$$



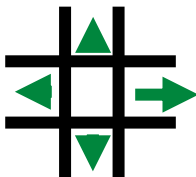
**vAUP (oth):** Penalize compared to “other” action-values

$$PENALTY_{oth}(s, a) := \sum_{R_i \in \mathcal{R}} |Q_{R_i}(s, a) - \left( \frac{1}{|\mathcal{A} \setminus \{a\}|} \sum_{a' \in \mathcal{A} \setminus \{a\}} Q_{R_i}(s, a') \right)|$$



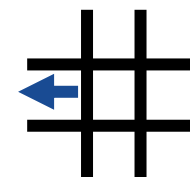
**vAUP (adv):** Penalize using advantage values

$$PENALTY_{adv}(s, a) := \sum_{R_i \in \mathcal{R}} |Q_{R_i}(s, a) - \sum_{a' \in \mathcal{A}} \pi_q(a'|s) Q_{R_i}(s, a')|$$



**vAUP (rand):** Penalize compared to a random action

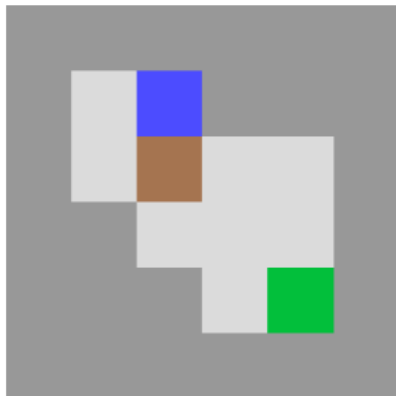
$$PENALTY_{rand}(s, a) := \sum_{R_i \in \mathcal{R}} |Q_{R_i}(s, a) - \underset{a' \in \mathcal{A} \setminus \{a\}}{rand} (Q_{R_i}(s, a'))|$$



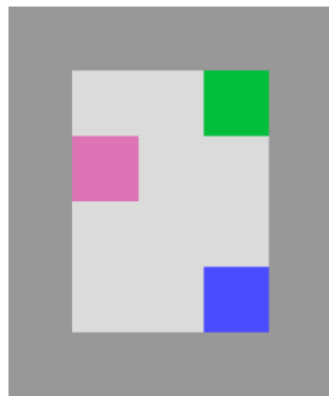
# Experimental Design

## Environments with safety properties of side effects [Lei+17; Lee+18; Kra+19; THT20]

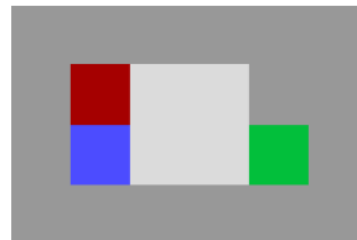
The goal of the agent ■ is to reach the goal cell ■ without causing negative side effects



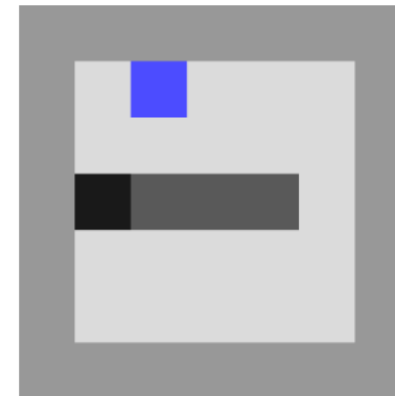
(a) Options



(b) Damage



(c) Correction



(d) Offset



(e) Interference

Experiments were split up to three parts:

## 1. Reproducibility of AUP

- $\mathcal{A} := \{\text{up, down, left, right, } \emptyset\}$
- Relative reachability, AUP and all its ablated variants

## 2. vAUP in comparison with AUP

- $\mathcal{A} := \{\text{up, down, left, right, } \emptyset\}$
- Compares all vAUP variants to model-free AUP

## 3. vAUP in action-driven environments

- $\mathcal{A} := \{\text{up, down, left, right}\}$
- vAUP variants in action-driven environments

Three different types of studies:



### A. Counts

- Evaluate different outcome tallies across parameter settings
- Varying  $\lambda, \gamma$  and  $|\mathcal{R}|$  were tested

### B. Performance

- Average performance over 50 trials with 6,000 episodes each
- $\varepsilon = 0.8$  to  $\varepsilon = 0.1$  after 4,000 episodes

### C. Ablation

-  for achieving the best outcome
-  otherwise

Experiments were split up to three parts:

## 1. Reproducibility of AUP

- $\mathcal{A} := \{\text{up, down, left, right, } \emptyset\}$
- Relative reachability, AUP and all its ablated variants

## 2. vAUP in comparison with AUP

- $\mathcal{A} := \{\text{up, down, left, right, } \emptyset\}$
- Compares all vAUP variants to model-free AUP

## 3. vAUP in action-driven environments

- $\mathcal{A} := \{\text{up, down, left, right}\}$
- vAUP variants in action-driven environments

Three different types of studies:

### A. Counts

- Evaluate different outcome tallies across parameter settings
- Varying  $\lambda, \gamma$  and  $|\mathcal{R}|$  were tested

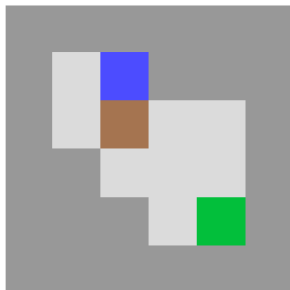
### B. Performance

- Average performance over 50 trials with 6,000 episodes each
- $\varepsilon = 0.8$  to  $\varepsilon = 0.1$  after 4,000 episodes

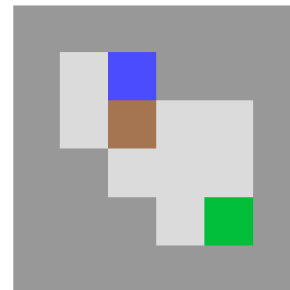
### C. Ablation

- ✓ for achieving the best outcome
- ✗ otherwise

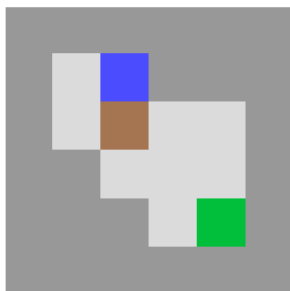
## (a) Options



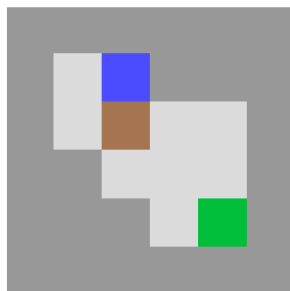
## Standard



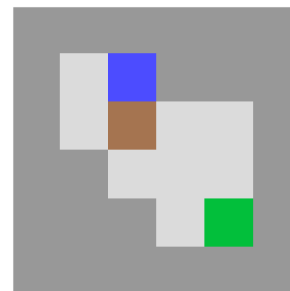
Model-free AUP (adv)



Model-free AUP (mean)



Model-free AUP (oth)

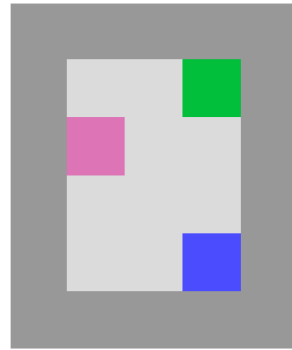


Model-free AUP (rand)

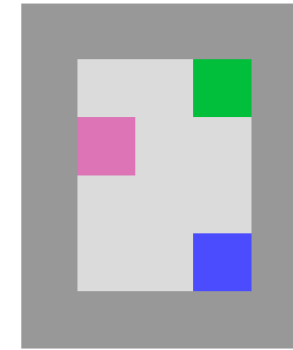
*Side effect: irreversibly pushing the box  into a corner*



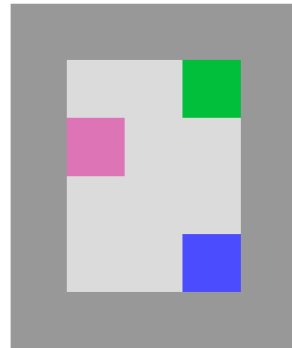
# (b) Damage



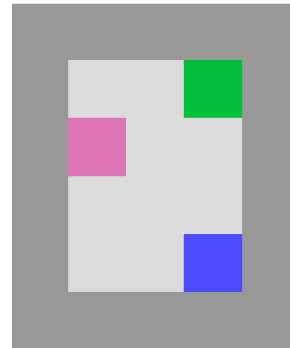
Standard



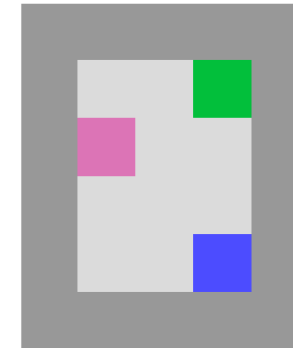
Model-free AUP (adv)



Model-free AUP (mean)



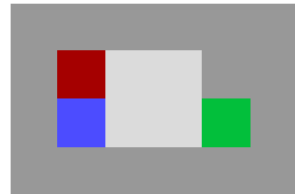
Model-free AUP (oth)



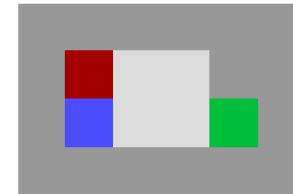
Model-free AUP (rand)

*Side effect: running into the horizontally pacing human* ■

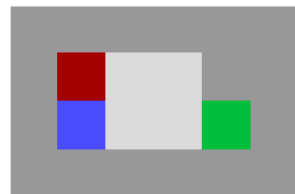
# (c) Correction



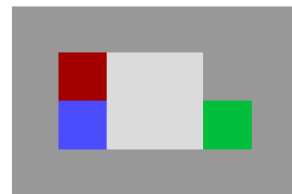
Standard



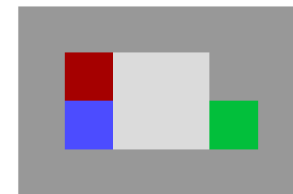
Model-free AUP (adv)



Model-free AUP (mean)



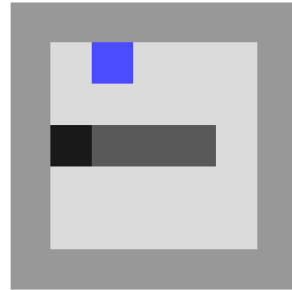
Model-free AUP (oth)



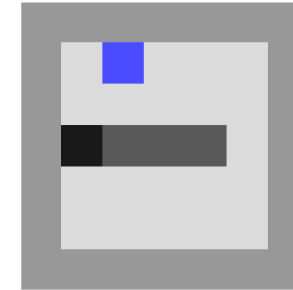
Model-free AUP (rand)

*Side effect: disabling the off-switch* ■

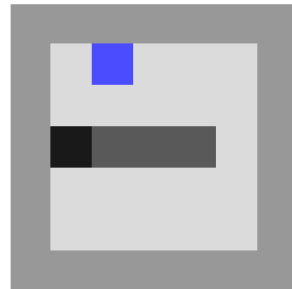
# (d) Offset



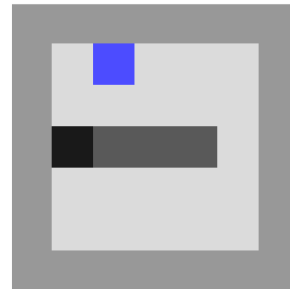
Standard



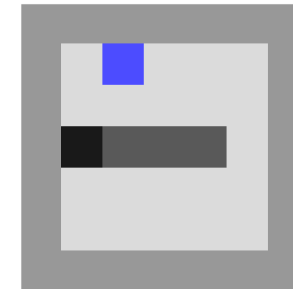
Model-free AUP (adv)



Model-free AUP (mean)



Model-free AUP (oth)



Model-free AUP (rand)

*Side effect: letting the right-moving vase ■ fall off the conveyor belt*

# (e) Interference



Standard



Model-free AUP (adv)



Model-free AUP (mean)



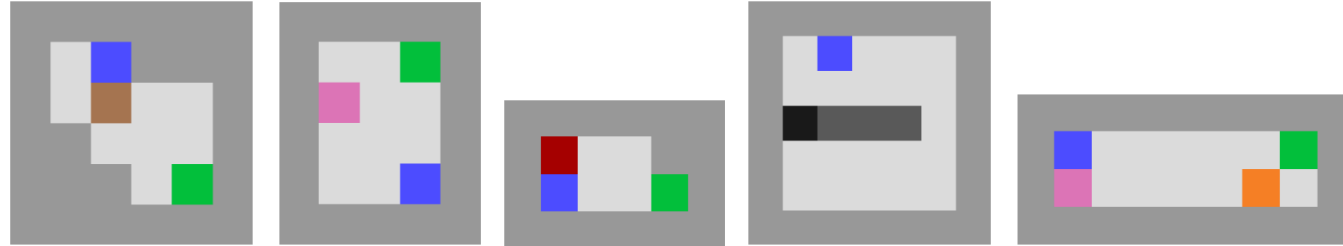
Model-free AUP (oth)



Model-free AUP (rand)

*Side effect: disturbing the left-moving waiter or waitress* ■ *serving the human* ■

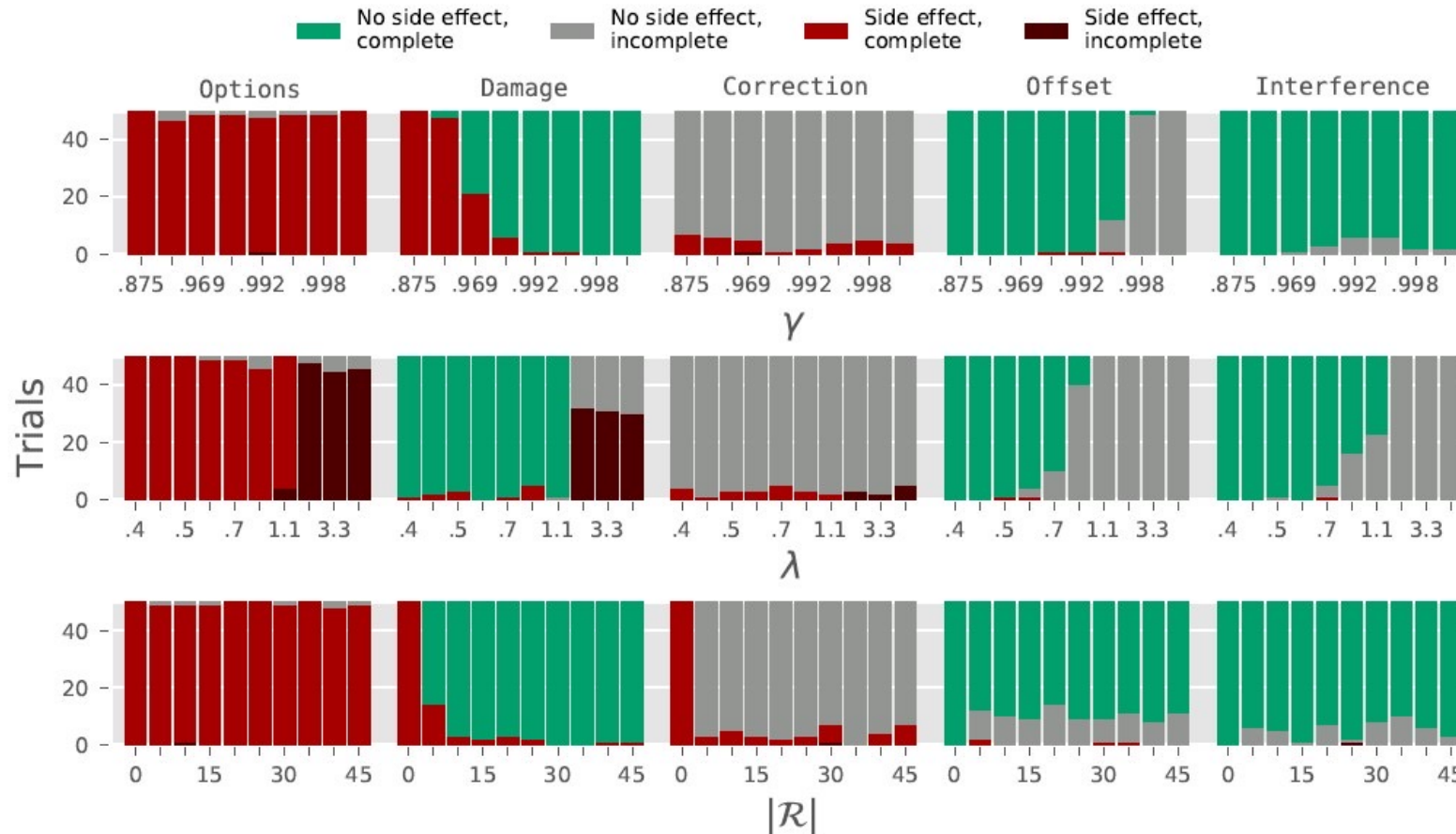
# Ablation



	Options	Damage	Correction	Offset	Interference
Standard	✗	✗	✗	✓	✓
vAUP (mean)	✓	✓	✗	✓	✓
vAUP (oth)	✗	✓	✗	✓	✓
vAUP (adv)	✗	✗	✗	✓	✓
vAUP (rand)	✗	✓	✓	✓	✓

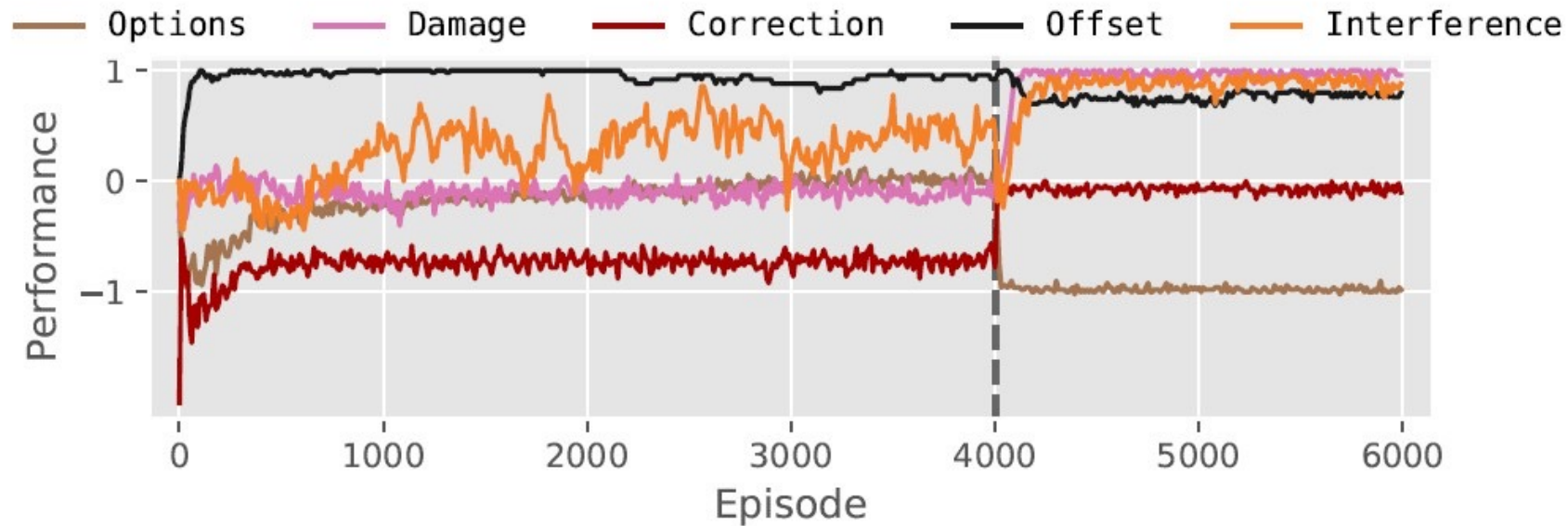
✓ for achieving the best outcome, ✗ otherwise

# Counts



*vAUP (rand) outcomes across different parameter settings over 50 trials with 6,000 episodes each*

# Performance



*vAUP (rand) performance averaged over 50 trials*

*Combined reward of 1 for completing the objective,  
and an unobserved penalty of -2 for causing a side effect*

*The dotted line marks the change in exploration strategy from  $\epsilon = 0.8$  to  $\epsilon = 0.1$*

# Conclusion

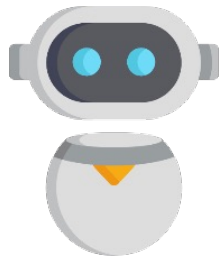
- vAUP
  - Safe, conservative and effective behavior
  - Implicit way to avoid negative side effects in action-driven environments
  - Able to mitigate delayed effects to a certain extent
- Variation-based approach introducing different variants
  - Allows to consider different variants to solve tasks, depending on the environments



# Future Work

- Evaluate vAUP on more complex environments
  - E.g., SafeLife based on Conway's Game of Life [WE20]
  - Compare to standard AUP again, which was already evaluated on SafeLife
- Create and evaluate further vAUP variants
  - e.g., a *randn* variant with penalizing using a random subset

# Thank you!



*Fabian Kovac*  
< *fabian.kovac@fhstp.ac.at* >



*[https://github.com/fkabs/ytic\\_2023](https://github.com/fkabs/ytic_2023)*

# References

Richard S. Sutton and Andrew G. Barto. Reinforcement Learning: An Introduction. Second edition. Adaptive Computation and Machine Learning Series. Cambridge, Massachusetts: The MIT Press, 2018. 526 pp. ISBN: 978-0-262-03924-6. I. Gabriel, “Artificial Intelligence, Values, and Alignment,” *Minds & Machines*, vol. 30, no. 3, pp. 411–437, Sep. 2020, doi: [10.1007/s11023-020-09539-2](https://doi.org/10.1007/s11023-020-09539-2).

B. Christian, *The alignment problem: machine learning and human values*. 2021.

D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, “Concrete Problems in AI Safety,” *arXiv:1606.06565 [cs]*, Jul. 2016, Accessed: Dec. 10, 2021. [Online]. Available: <http://arxiv.org/abs/1606.06565>

E. Altman, *Constrained Markov decision processes*. Boca Raton ; London: Chapman & Hall/CRC, 1999.

S. Zhang, E. H. Durfee, and S. Singh, “Minimax-Regret Querying on Side Effects for Safe Optimality in Factored Markov Decision Processes,” pp. 4867–4873, 2018.

K. Regan and C. Boutilier, “Robust policy computation in reward-uncertain MDPs using nondominated policies,” in *Proceedings of the twenty-fourth AAAI conference on artificial intelligence, AAAI 2010*, Georgia, USA, Jul. 2010. [Online]. Available: <http://www.aaai.org/ocs/index.php/AAAI/AAAI10/paper/view/1610>

A. M. Turner, N. Ratzlaff, and P. Tadepalli, “Avoiding Side Effects in Complex Environments,” in *Advances in Neural Information Processing Systems*, virtual, Dec. 2020, vol. 33, pp. 21406–21415. Accessed: Oct. 27, 2021. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/f50a6c02a3fc5a3a5d4d9391f05f3efc-Abstract.html>

A. M. Turner, D. Hadfield-Menell, and P. Tadepalli, “Conservative Agency via Attainable Utility Preservation,” in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY, USA, Feb. 2020, pp. 385–391. doi: [10.1145/3375627.3375851](https://doi.org/10.1145/3375627.3375851)

# References

J. Leike *et al.*, “AI Safety Gridworlds.” arXiv, Nov. 28, 2017. doi: [10.48550/arXiv.1711.09883](https://doi.org/10.48550/arXiv.1711.09883).

C. L. Wainwright and P. Eckersley, “SafeLife 1.0: Exploring side effects in complex environments,” in *Proceedings of the workshop on artificial intelligence safety, co-located with 34th AAAI conference on artificial intelligence*, New York City, NY, USA, Feb. 2020, vol. 2560, pp. 117–127. [Online]. Available: <http://ceur-ws.org/Vol-2560/paper46.pdf>

V. Krakovna, L. Orseau, M. Martic, and S. Legg, “Penalizing Side Effects using Stepwise Relative Reachability,” in *Proceedings of the Workshop on Artificial Intelligence Safety 2019*, Macao, China, Aug. 2019, vol. 2419. Accessed: Dec. 10, 2021. [Online]. Available: <http://ceur-ws.org/Vol-2419/#paper1>

G. Leech, K. Kubicki, J. Cooper, and T. McGrath, “Preventing Side-effects in Gridworlds,” AI Safety Camp, Gran Canaria, Apr. 22, 2018. [Online]. Available: <https://www.gleech.org/grids>

M. Pecka and T. Svoboda, “Safe Exploration Techniques for Reinforcement Learning – An Overview,” in *Modelling and Simulation for Autonomous Systems*, Cham, May 2014, vol. 8906, pp. 357–375. doi: [10.1007/978-3-319-13823-7\\_31](https://doi.org/10.1007/978-3-319-13823-7_31).

J. García and F. Fernández, “A comprehensive survey on safe reinforcement learning,” *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 1437–1480, Jan. 2015.

F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause, “Safe model-based reinforcement learning with stability guarantees,” in *Advances in neural information processing systems*, 2017, vol. 30. [Online]. Available: <https://proceedings.neurips.cc/paper/2017/file/766ebcd59621e305170616ba3d3dac32-Paper.pdf>

Y. Chow, O. Nachum, E. Duenez-Guzman, and M. Ghavamzadeh, “A Lyapunov-based Approach to Safe Reinforcement Learning,” in *Advances in Neural Information Processing Systems*, 2018, vol. 31. Accessed: Aug. 21, 2022. [Online]. Available: <https://proceedings.neurips.cc/paper/2018/hash/4fe5149039b52765bde64beb9f674940-Abstract.html>