Reinforcement Learning from Human Feedback for Cyber-Physical Systems

On the Potential of Self-Supervised Pretraining

Timo Kaufmann, Viktor Bengs, Eyke Hüllermeier
LMU Munich

Hamburg, 29.03.2023

Bringing Reinforcement Learning into the Real World

Part of the ONE Munich Strategy Forum Project:

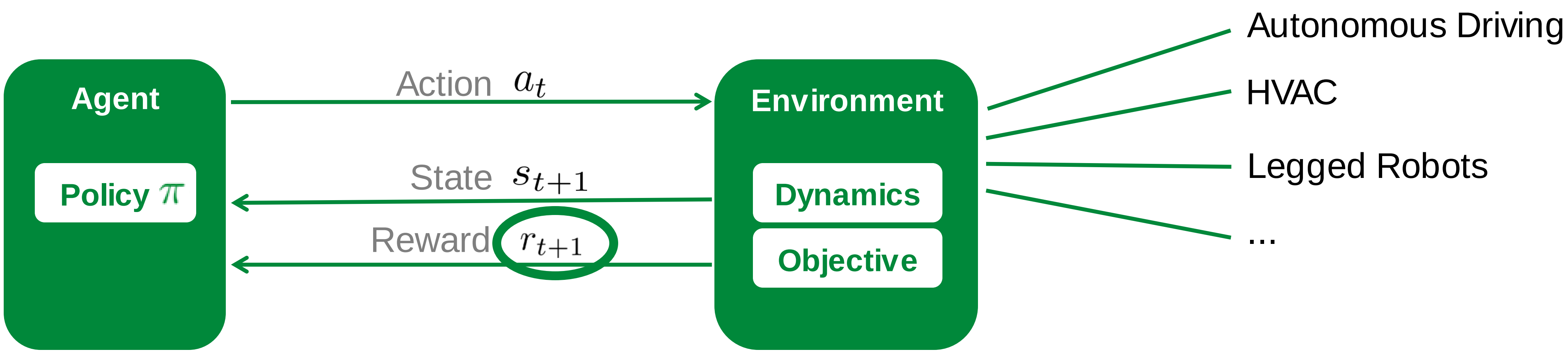
Next generation Human-Centered Robotics

Image: Midjourney

https://human-centered-robotics.de

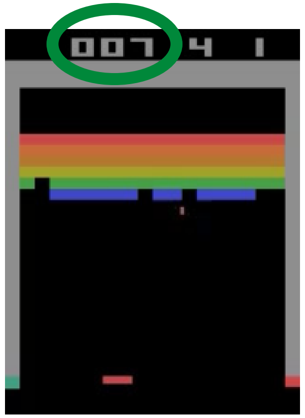
Reinforcement Learning
from
Human Feedback

Reinforcement Learning for Cyber-Physical Systems



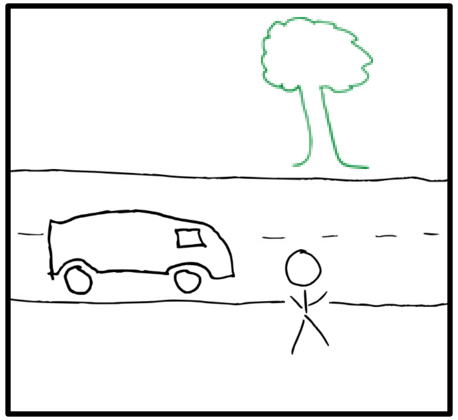
Objective: $\max \sum_t \gamma^t r_t$

Reinforcement Learning Favors Quantifiable Tasks

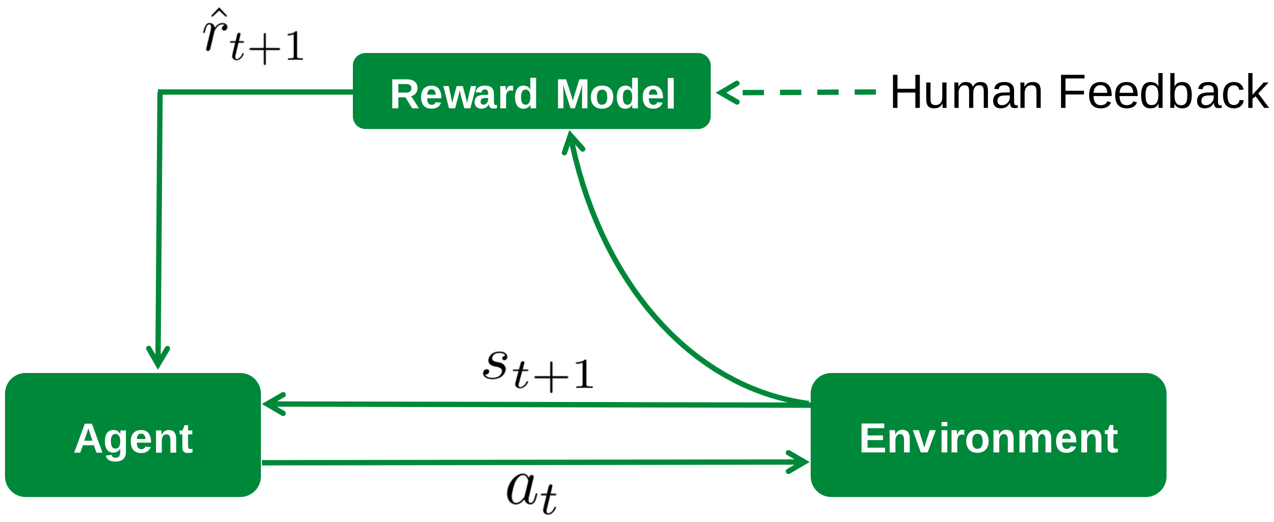


$$r_{t+1} = R\left(s_t, \underbrace{\text{left}}_{a_t}\right) = cur\_score - prev\_score$$

The Limits of Classical Reinforcement Learning

$R\Big($  $, a_t\Big) = ?$

Reinforcement Learning from Human Feedback with Reward Modelling



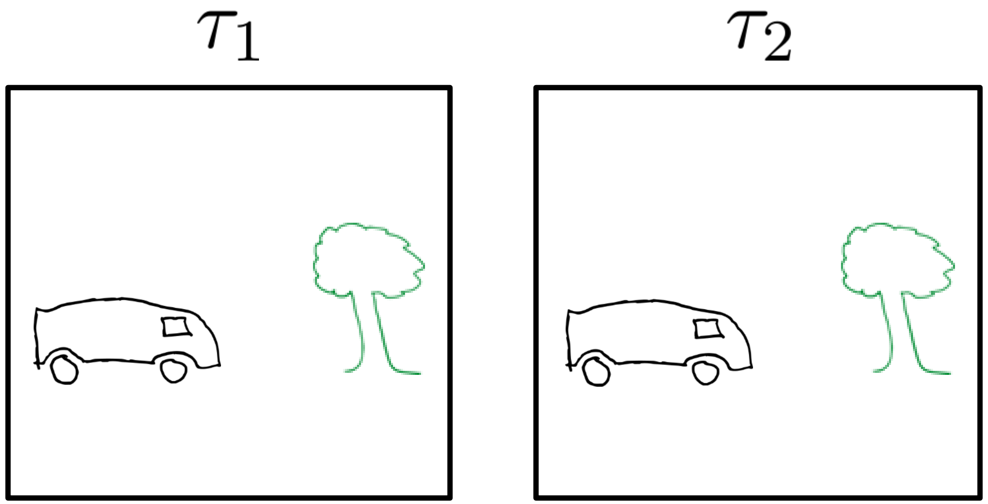
Goal: $\tau_1 \succ \tau_2$

Where: $\tau_i = (s_1^i, a_1^i, s_2^i, a_2^i, \ldots, s_n^i, a_n^i)$

Proxy Objective: $\max \sum_t \gamma^t \hat{r}_t$

Reproduced from Christiano et al., 2017

Rewards from Pairwise Trajectory Preferences

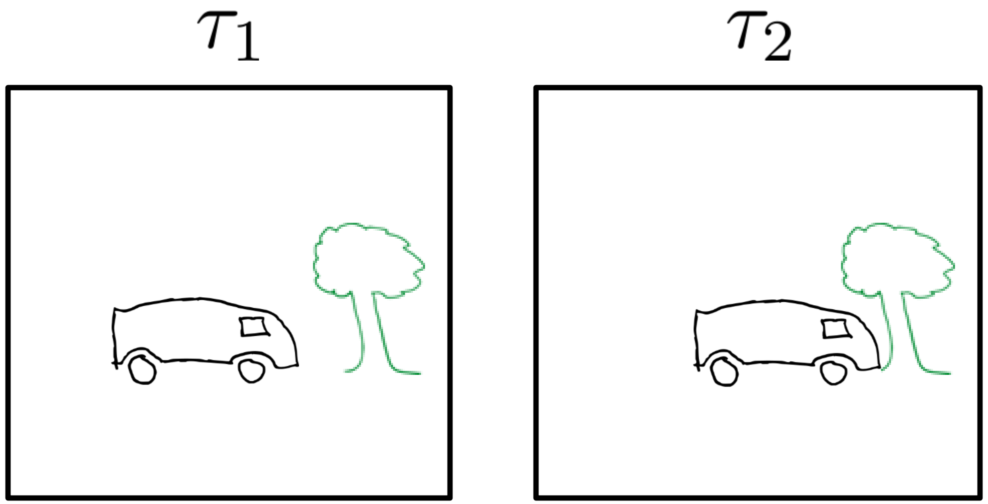


$$\tau_1 \succ \tau_2 \Rightarrow R(\tau_1) > R(\tau_2)$$

Bradley-Terry links preferences to rewards:

$$P[\tau_1 \succ \tau_2] = \mathrm{sofmax}_1\big(R(\tau_1), R(\tau_2)\big)$$

Rewards from Pairwise Trajectory Preferences



$$\tau_1 \succ \tau_2 \Rightarrow R(\tau_1) > R(\tau_2)$$

Bradley-Terry links preferences to rewards:

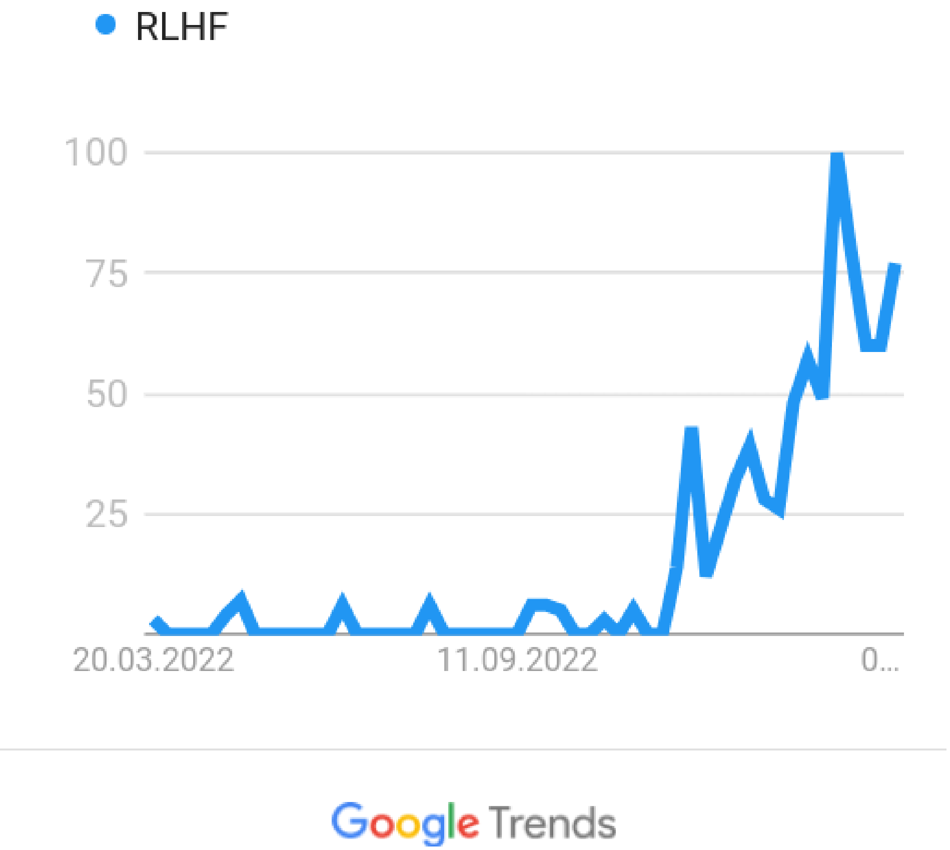
$$P[\tau_1 \succ \tau_2] = \mathrm{sofmax}_1\big(R(\tau_1), R(\tau_2)\big)$$

Past and Present of RLHF

- Emerged from preference-based RL.
  Cheng et al., 2011; Akrour et al., 2011

- RL for fine-tuning foundation models:
  ChatGPT, GPT4.

- Increasing relevance due to use of RL in the real world.



Compare Ouyang et al., 2022

Self-Supervised Pretraining

Self-Supervised Learning

Learn without explicit supervision!

Predictive

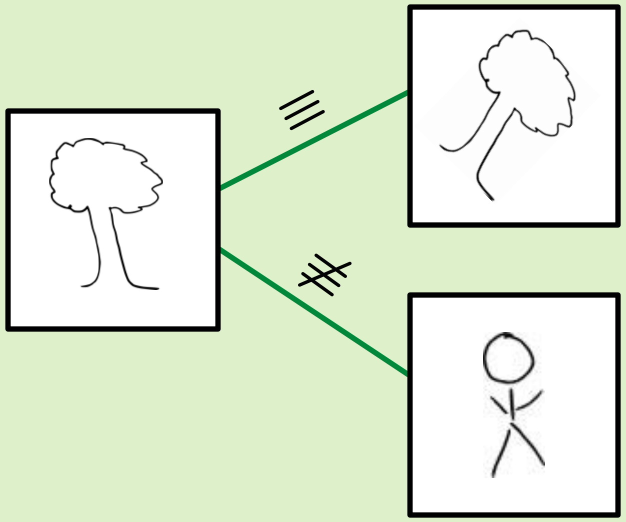
tree

The squirrel is climbing the _____

skyscraper

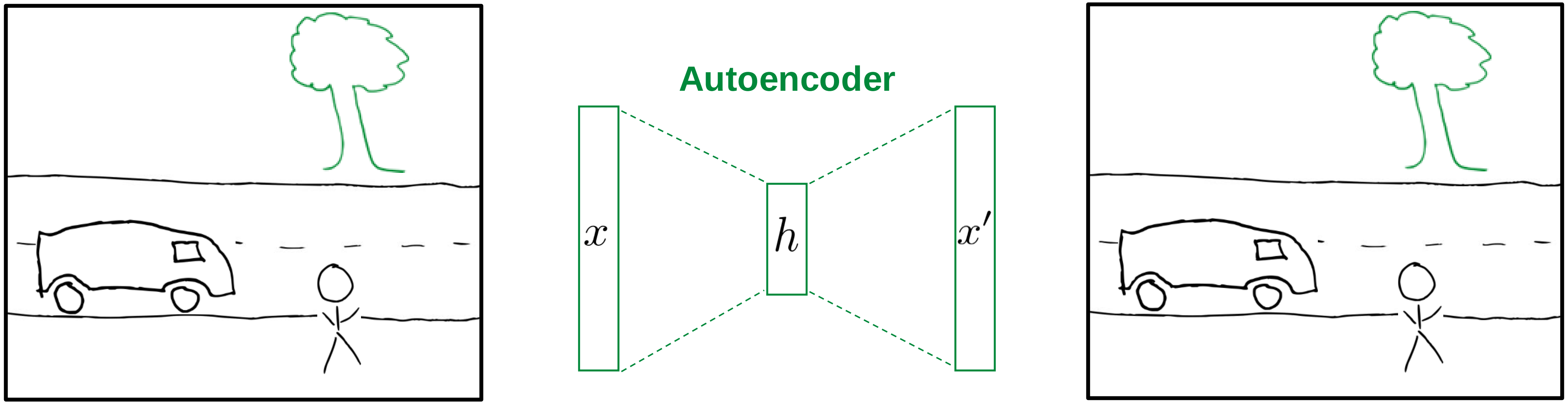
Challenge: Represent distribution.

Contrastive

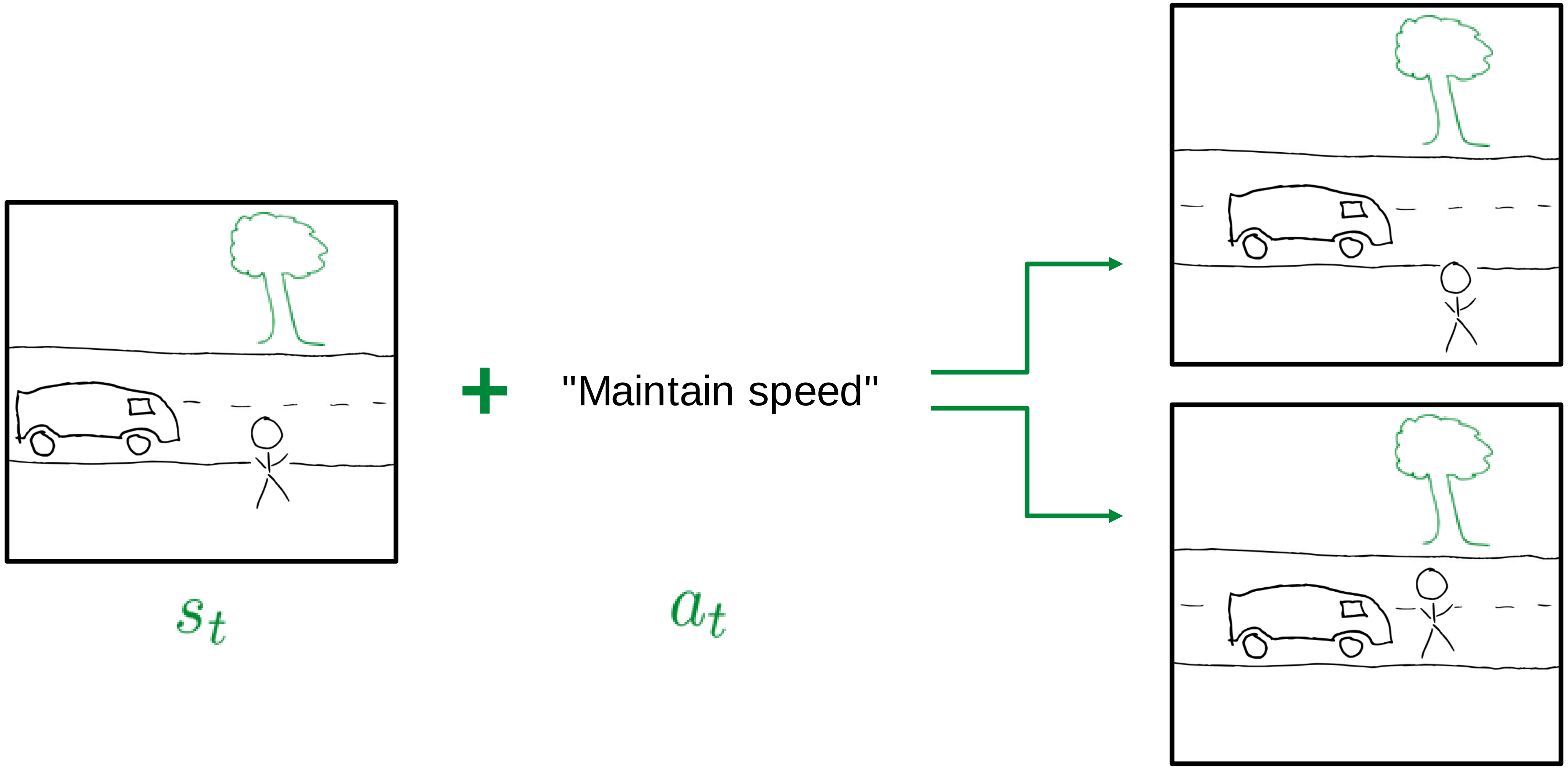


Challenge: Find hard negatives.

Self-Supervised State Representation Learning



Self-Supervised World Model Learning

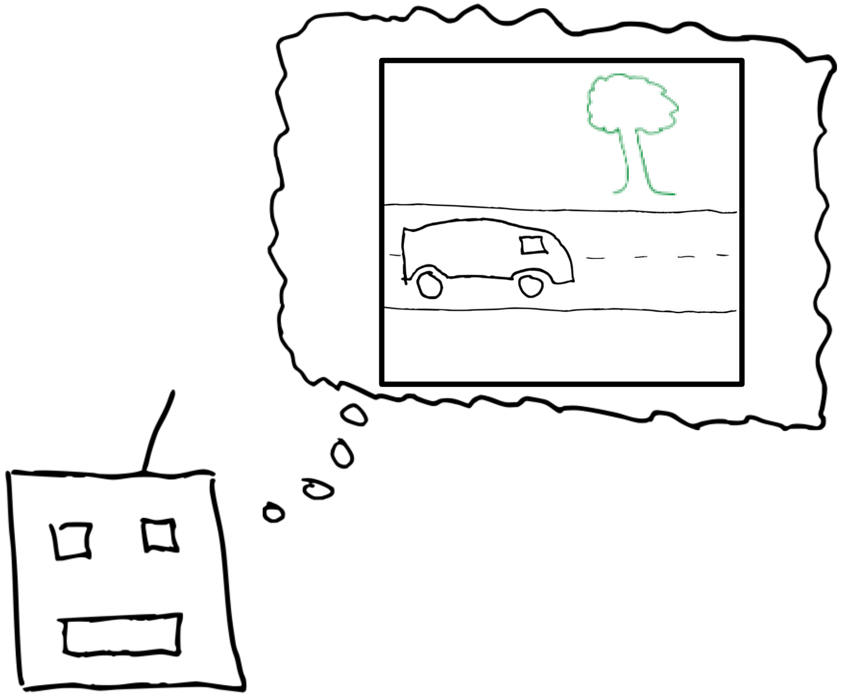


s_t a_t

$$\hat{s}_{t+1} \sim p_\theta(\hat{s}_{t+1} \mid s_t, a_t)$$

Trees stay in place, humans may move.

World Models Enable Query Synthesis



"Imagination": Repeatedly sample

$$\hat{s}_{t+1} \sim p_\theta(\hat{s}_{t+1} \mid s_t, a_t)$$
$$a_{t+1} \sim \pi(\hat{s}_{t+1})$$

Benefits of Pretraining for RLHF

Sample Efficiency | Transfer | Safety | Robustness | Reward Exploration

Sample Efficiency for Preference Learning

$$\begin{array}{cl} \left[\begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_{65536} \end{array}\right] & \begin{array}{l} \text{Red channel of first pixel} \\ \text{Green channel of first pixel} \\ \\ \text{Blue channel of last pixel} \end{array} \end{array}$$

Many noisy dimensions

$$\begin{array}{cl} \left[\begin{array}{c} z_1 \\ z_2 \\ \vdots \\ z_{64} \end{array}\right] & \begin{array}{l} \text{Distance to pedestrian} \\ \text{Light conditions} \\ \\ \text{Speed limit} \end{array} \end{array}$$

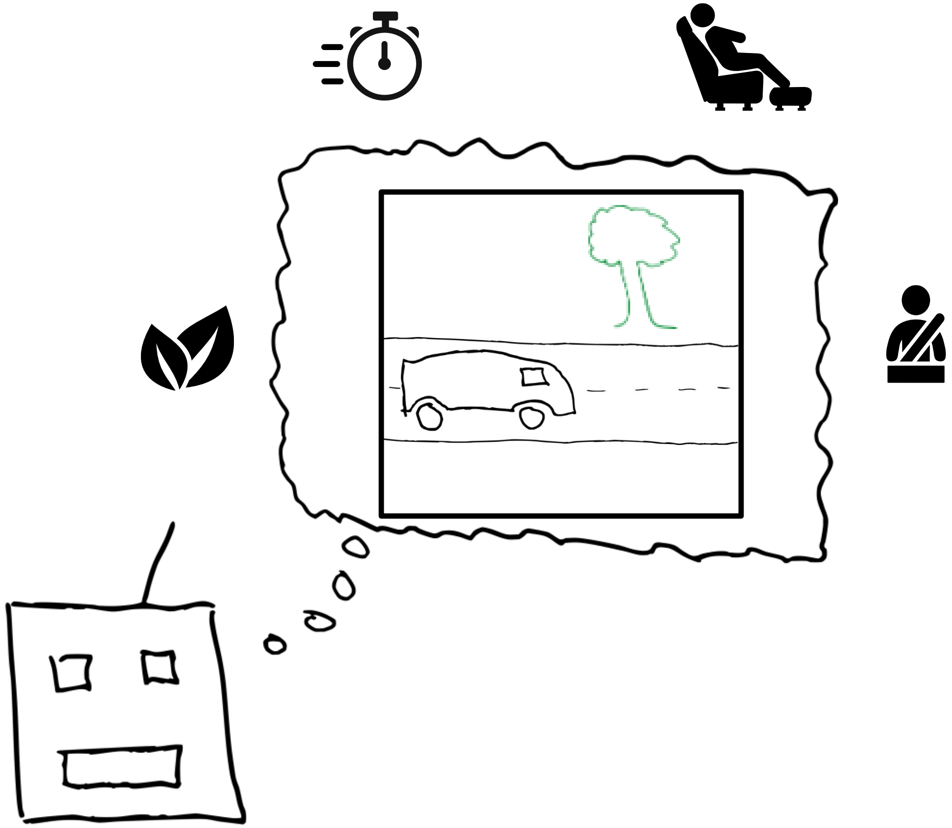
Few highly informative dimensions

Learn concepts first, then preferences.

Challenge: Auxiliary task design.

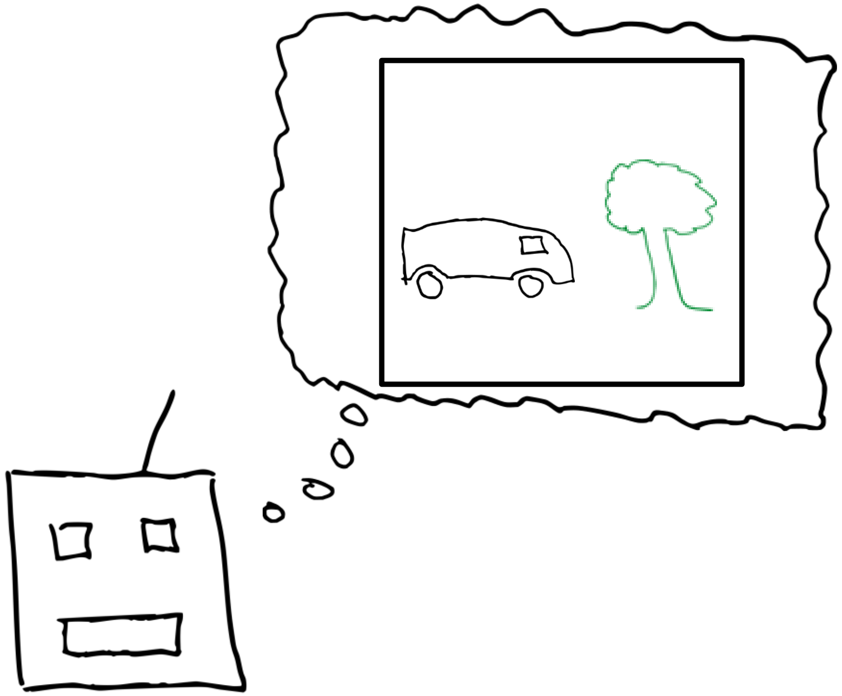
Transfer Enabled by Representations

- Representations can be task-independent.
- Can reuse representations for faster adaptation.
- Can scale training over multiple tasks.
- Potentially even adapt entirely in imagination.

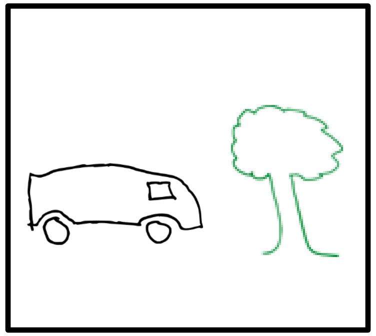


Icons: uxwing.com

Safety Through Query Synthesis



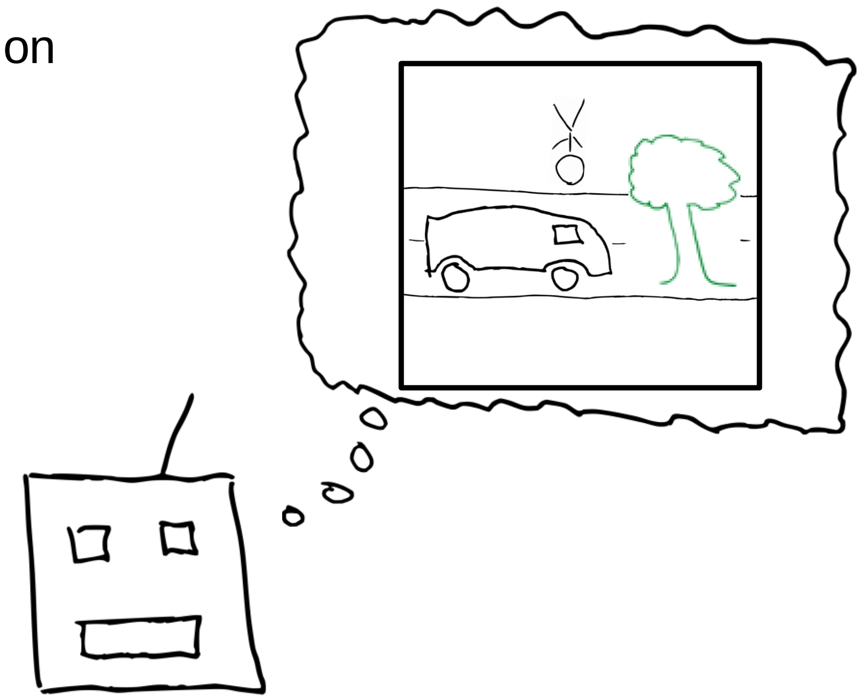
Instead of



Imagined crashes do not hurt.

Robustness Through Query Synthesis

- Synthesis enables feedback on rare events, outliers and uncertain regions.



Reward Exploration with World Models

State Space Exploration

- Challenge in RL: Exploration / exploitation tradeoff.
- Exploration is commonly incentivized with intrinsic motivation.

$$r_t = r_t^{\text{task}} + r_t^{\text{intrinsic}}$$

- Example: Reward based on estimated state novelty.
- Problem: The policy is optimized to seek states that were previously novel – but are not anymore! Chases an outdated concept of novelty.
- Possible solution: Optimize exploration policy "in imagination", deploy "in real" (Plan2Explore).

Reward Space Exploration

- In RLHF additionally: Reward exploration!
- Similar techniques can be used.

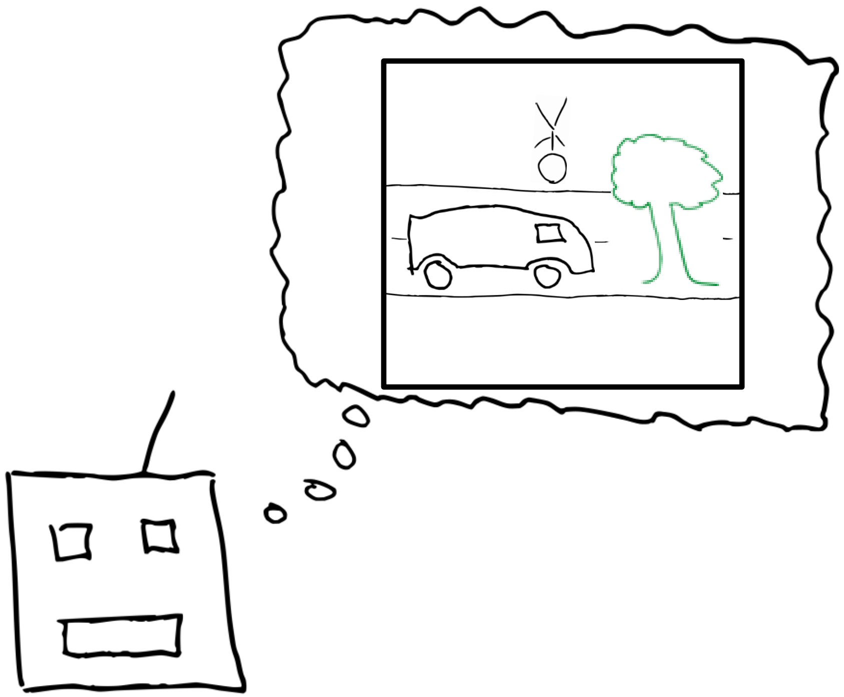
$$r_t = \hat{r}_t^{\text{task}} + r_t^{\text{intrinsic}}$$

- Reward uncertainty in the reward model.
- Plan2Explore approach may be used here!

Sekar et al., 2020: Planning to Explore via Self-Supervised World Models

Conclusion

- RL can enable new use-cases for CPS.
- Human feedback is crucial to make this practical.
- Self-supervised pretraining helps with
  - sample-efficiency,
  - transfer learning,
  - safety,
  - robustness and
  - reward exploration.



timokaufmann.com @timokauf

Questions?

References

- [Christiano et al., 2017]: Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep Reinforcement Learning from Human Preferences. Advances in Neural Information Processing Systems.
- [Ouyang et al., 2022]: Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Gray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems.
- [Sekar et al., 2020]: Sekar, R., Rybkin, O., Daniilidis, K., Abbeel, P., Hafner, D., & Pathak, D. (2020). Planning to Explore via Self-Supervised World Models. Proceedings of the 37th International Conference on Machine Learning.

- [Kaufmann et al., 2023]: Kaufmann, T., Bengs, V., & Hüllermeier, E. (2023). Reinforcement Learning from Human Feedback for Cyber-Physical Systems: On the Potential of Self-Supervised Pretraining.