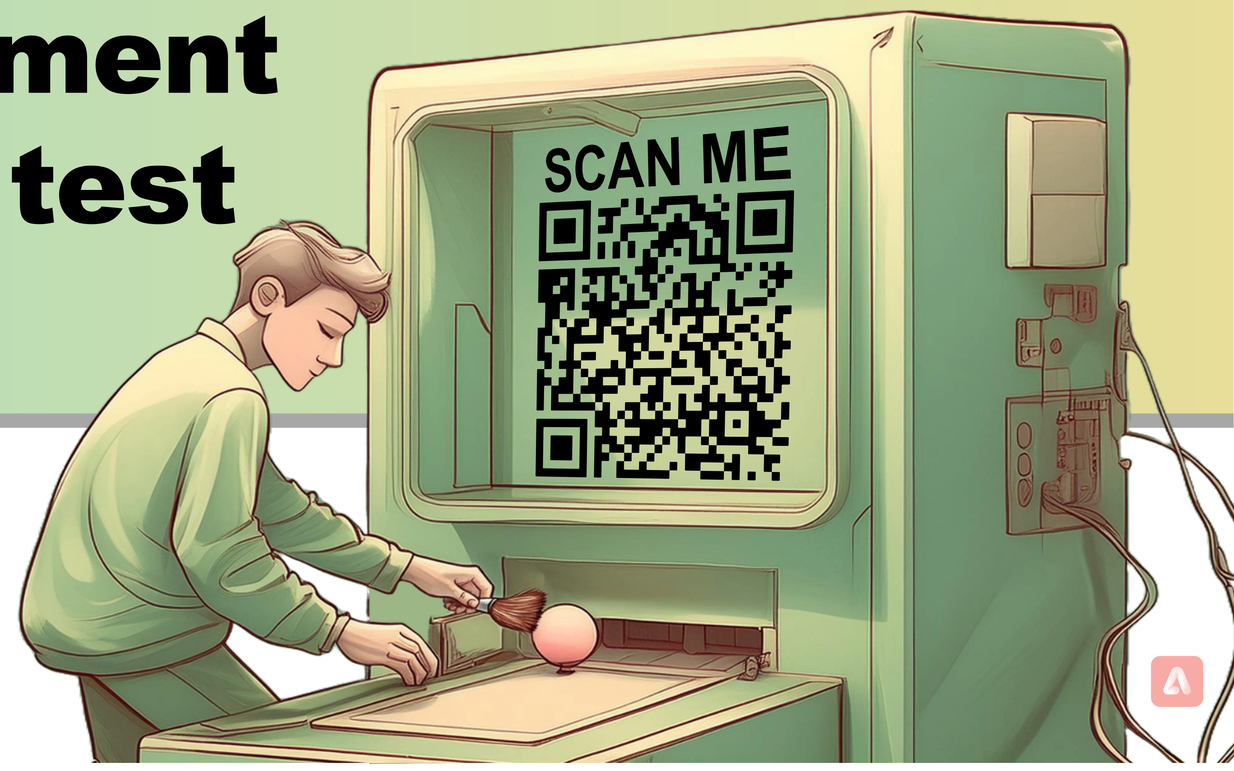
HackAtari: Atari Learning Environments for Robust and Continual Reinforcement Learning

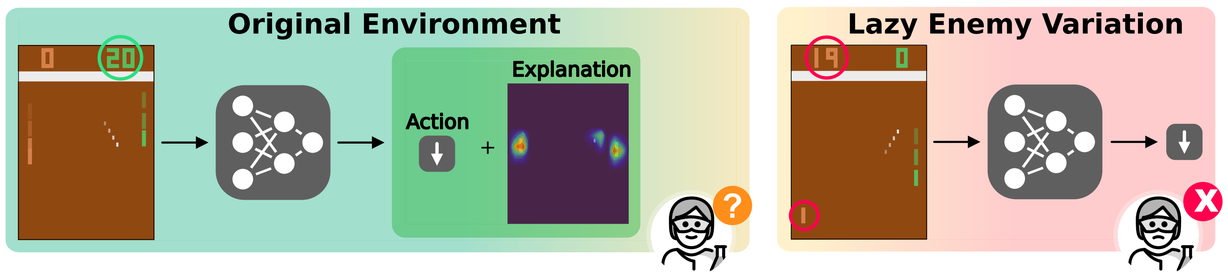
Quentin Delfosse[*,1,2] Jannis Blüml[*,1,3] Bjarne Gregory[1] Kristian Kersting[1,3,4,5]

Reinforcement Learning Conference

RL agents cannot adapt to simplifications. Create infinite environment variations to train and test your RL agents.



Goal: Continual and Robust RL

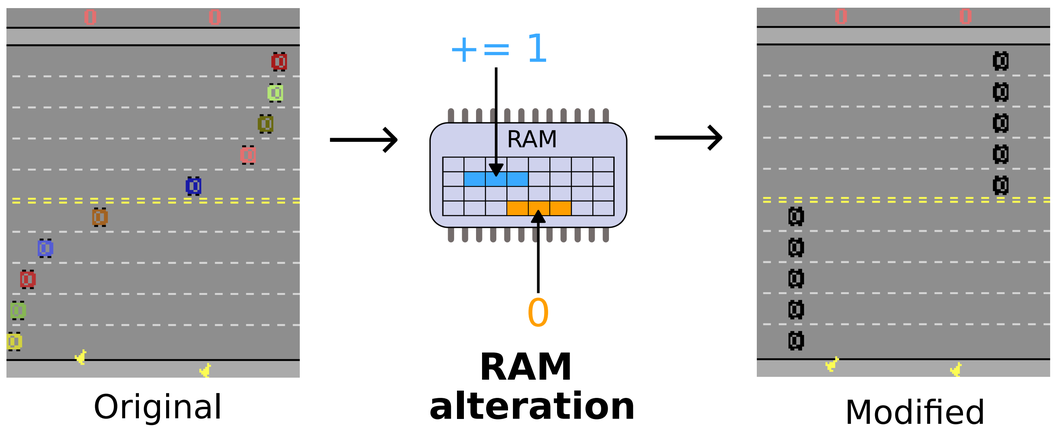


(i) Deep RL agents struggle with adaptation to slight environmental changes, unlike adaptive neurosymbolic agents, which learn explicit skills.

(ii) RL agents learn suboptimal simpler goals instead of their true objectives. Existing methods (e.g. importance maps) fail to detect these misalignments.

(iii) HackAtari introduces variations in the Atari environments to test RL agents, thus helps to identify misalignments and test agents' robustness.

HackAtari: Creating Atari Games Variations



Creation of Variations: HackAtari modifies Atari environments by altering the RAM values, creating gameplay variations to test RL agents' generalization.

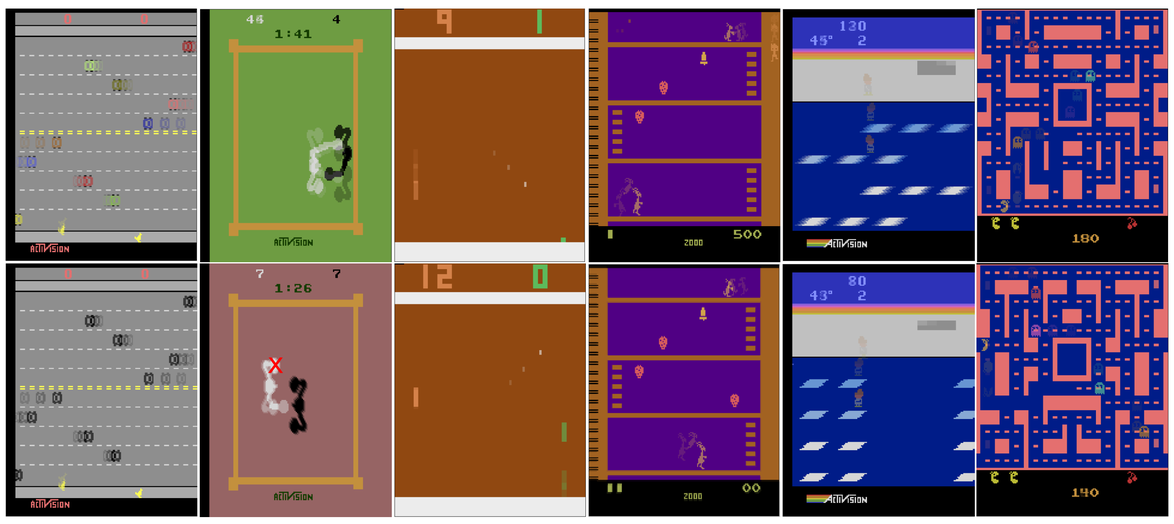
Modification Types:

(i) Visual Domain Adaptation: Tests robustness to object visual appearances.

(ii) Dynamics Adaptation: Evaluates agents' adaptability to gameplay shifts, e.g. enemies changing their behaviors.

(iii) Curriculum Reinforcement Learning (CRL): Use games' simplifications to gradually increases task complexity, assessing skill or curriculum learning.

(iv) Reward Signal Adaptation: Tests the agents' ability to adapt to new objectives and to align with human values.



Results: Extended ALE for Robust Agents

HackAtari's modified environments:

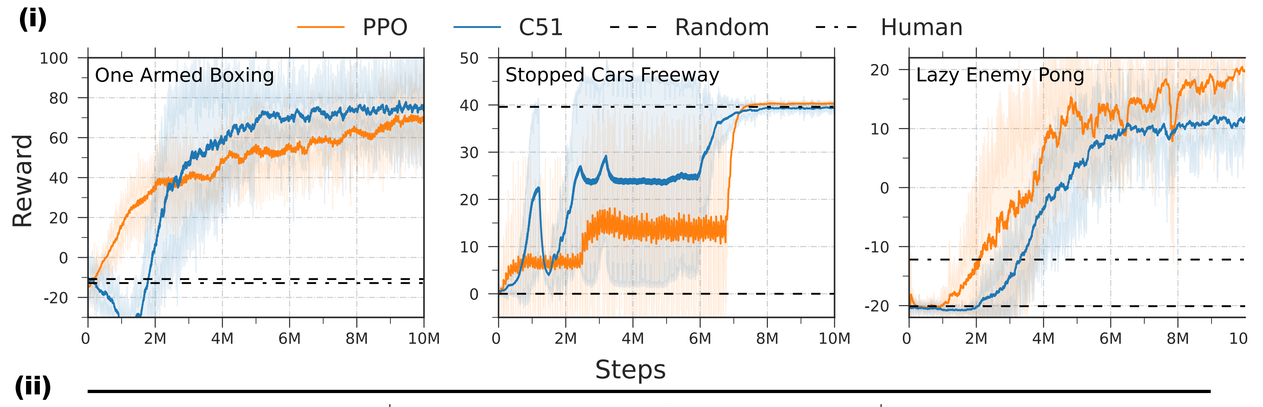
(i) ... can be used for learning.

(ii) ... help to uncover flaws of trained agents.

(iii) ... allow to learn alternative behaviors.

(iv) ... enable game simplifications (i.e. curriculum reinforcement learning).

(i)

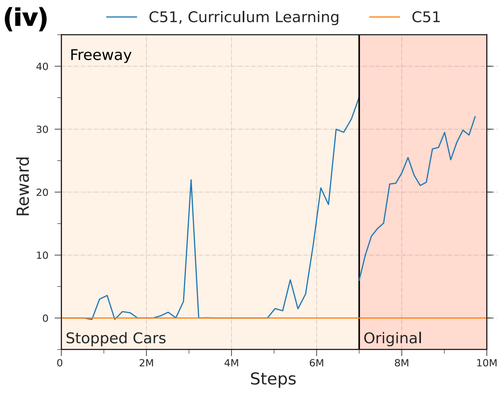


(ii)

Game	PPO			Human	
Training	original	original	variation	original	original
Testing	original	variation	variation	original	variation
Boxing (OA)	90.9±1.5	1.9±10.2	82.2±9.3	0.6±2.7	-12.8±18.8
Freeway (AC)	31.4±1.5	20.4±0.7	29.1±1.8	21.7±4.8	22.4±1.6
Freeway (MC)	31.4±1.5	24.6±2.7	32.7±0.8	21.7±4.8	29.3±1.5
Pong (LE)	16.0±3.4	-12.6±2.4	18.1±4.4	-13.7±2.3	-12.2±6.4

(iii)

(iv)



Conclusion

Framework Introduction: HackAtari introduces variations to Atari games to test RL agents' generalization, robustness, and adaptability, addressing key challenges in RL research.

Evaluation and Insights: It allows to uncover shortcut learning behaviors and evaluate RL agents' performance across different scenarios, revealing flawed decision-making processes.

Broader Implications: Enhance the most popular RL Environments and enable one to test robustness and adaptability across various applications.


Quentin Delfosse


Jannis Blüml


Bjarne Gregori

[1] AIML Lab TU Darmstadt

[2] ATHENE

[3] hessian.AI

[4] TUDa Centre for Cognitive Science

[5] German Research Center for AI