Deep Reinforcement Learning Agents are not even close to Human Intelligence

Quentin Delfosse, Jannis Blüml, Fabian Tatai, Théo Vincent, Bjarne Gregori, Elisabeth Dillies, Jan Peters, Constantin Rothkopf, Kristian Kersting

EWRL18

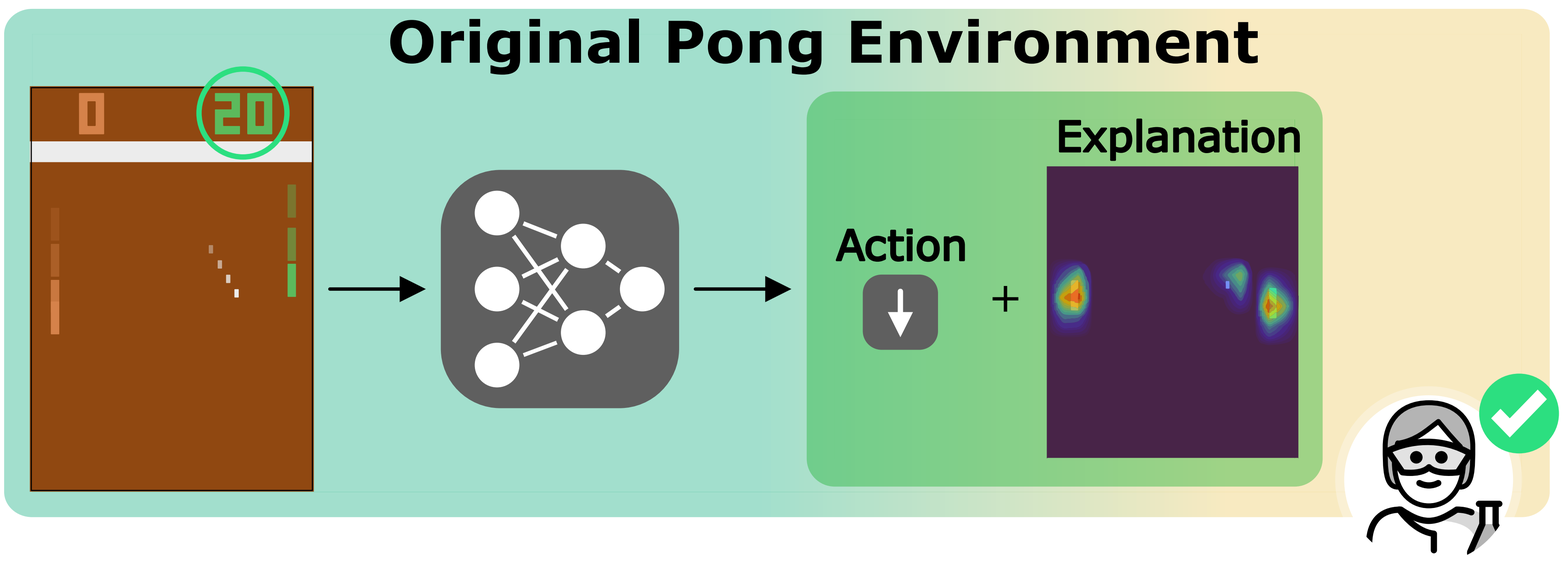
Paper & Code

RL agents learn shortcuts! They cannot adapt to task simplifications.

Problem

(i) **Deep RL agents struggle to adapt** even to slight environmental changes, like freezing the enemy in Pong.

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

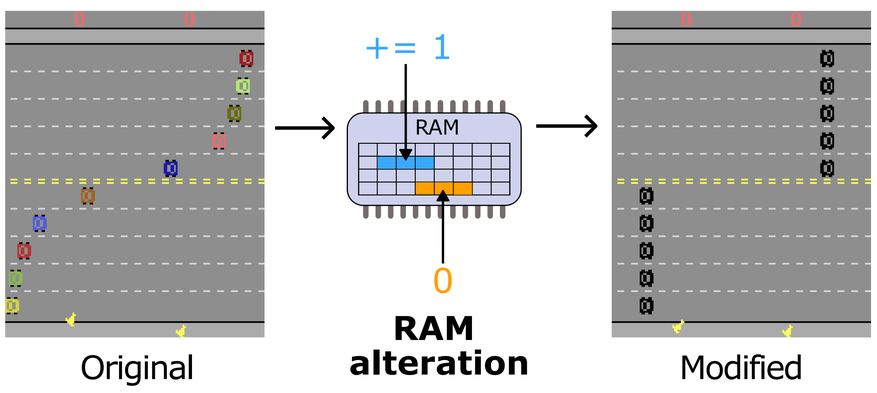


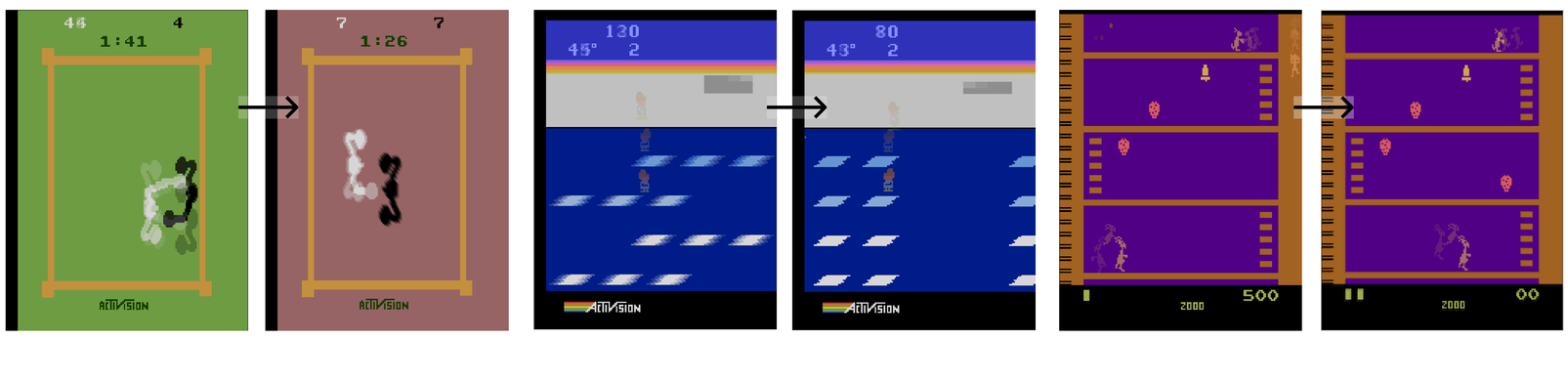
Evaluating on training environment leads to perfect score ✅, consistent actions ✅, and intuitive explanation maps ✅.

However, changing the enemy's behavior prevent the agent from catching the ball.

HackAtari

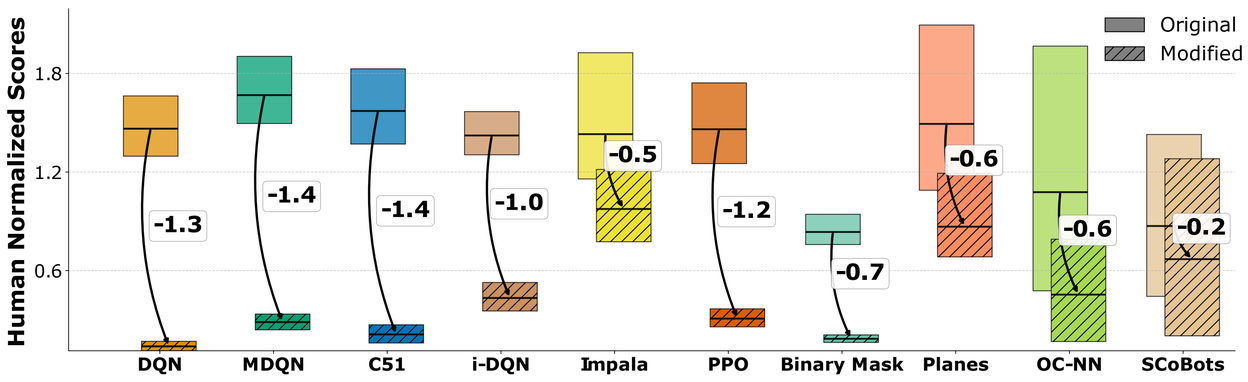
HackAtari introduces **variations in ALE games**. You can use it to detect that **your RL agents are misaligned**.



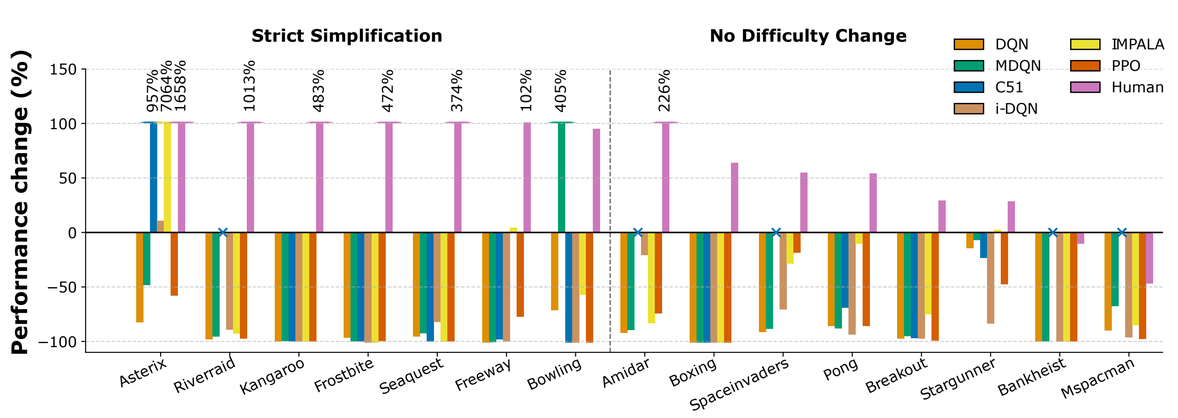


Results

Using HackAtati, we show that **DRL agents,** contrary to humans**, fail to adapt to tasks simplifications.**



Average performances drop of RL algorithms on tasks simplifications.



Detailed (per-game) performance changes of RL agents and humans.

Quentin Delfosse
Jannis Blüml
Théo Vincent

OCAtari HackAtari JAXAtari

AIML Lab TU Darmstadt
ATHENE
hessian.AI
IAS Lab TU Darmstadt
TUDa Centre for Cognitive Science
Sorbonne Université
German Research Center for AI