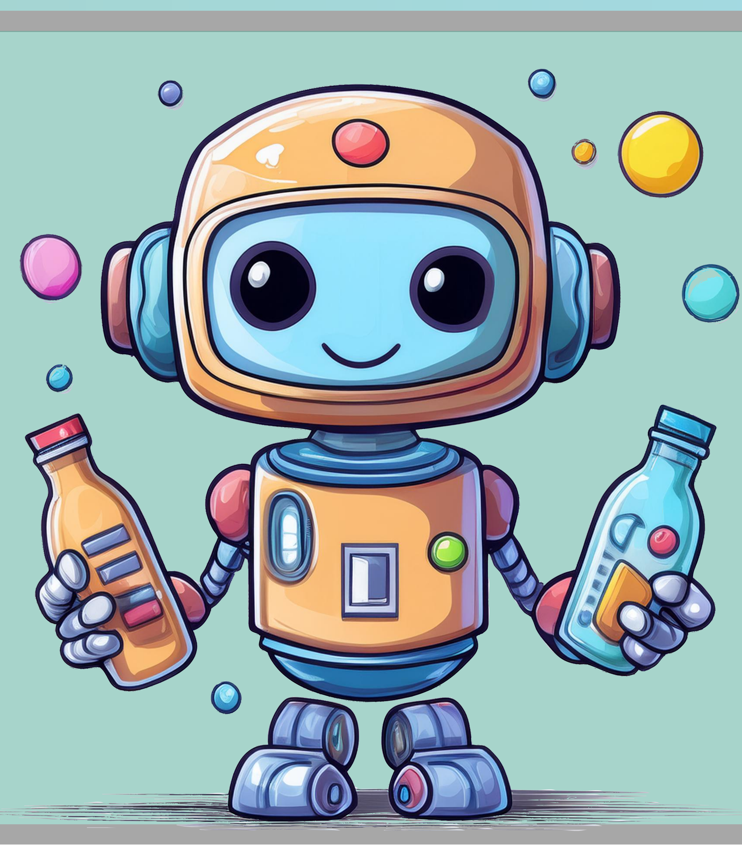
Interpretable Concept Bottlenecks to Align Reinforcement Learning Agents

Quentin Delfosse[1,2] Sebastian Sztwiertnia[1] Mark Rothermel[1] Wolfgang Stammer[1,4] Kristian Kersting[1,3,5]

Reinforcement Learning Conference

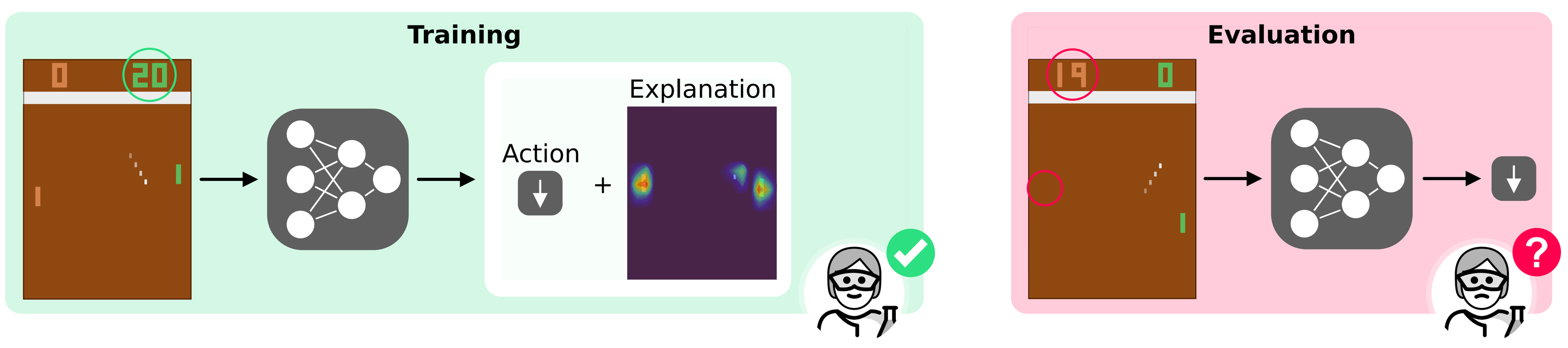
Deep agents perform undetectable Shortcut Reinforcement Learning. RL agents must utilize human understandable concepts.



SCAN ME

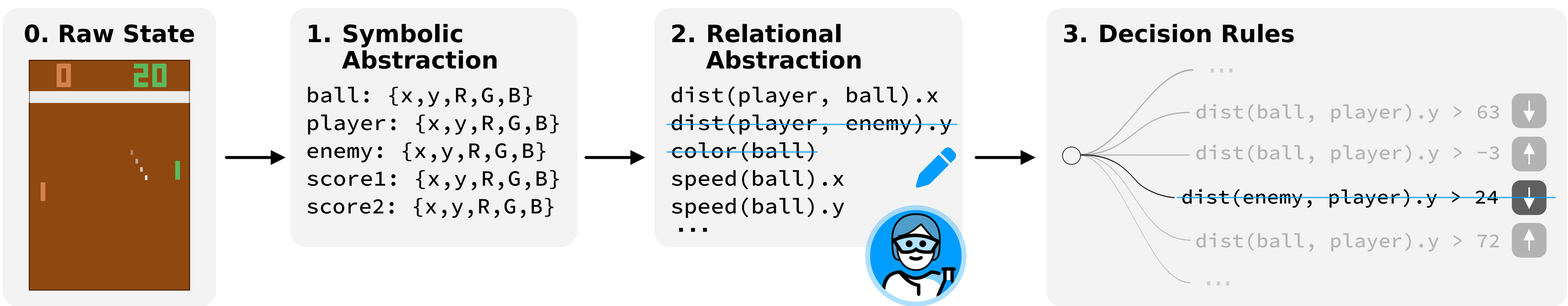
Goal: Interpretable RL agents

Deep RL agents inspection falls short



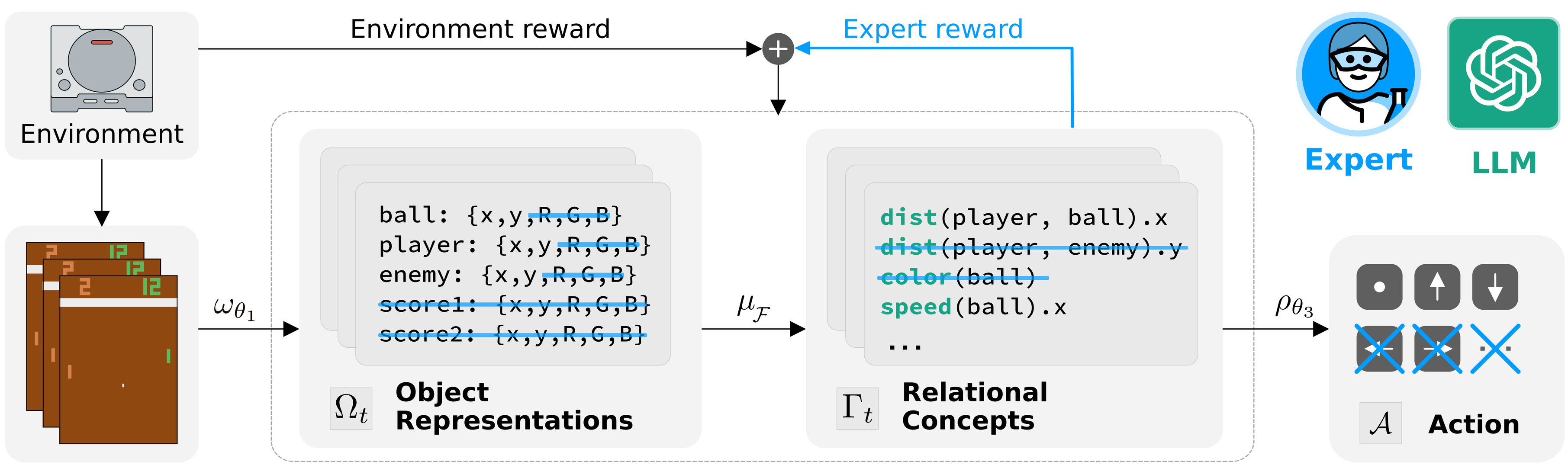
Deep RL agents learn hidden shortcut within **misaligned policies,** that **fail to generalize** to simpler scenarios ...

SCoBots' inspection and revision mitigate misalignment



... contrary to **Successive Concept Bottlenecks Agents** (SCoBots), that allow for simple **inspection** and **revision.**

SCoBots: Interpretable Concept Bottlenecks



We create **inspectable and guidable RL agents,** with **successive concept bottlenecks (CB)** as a policy:

(i) Object detection CB: Extracts **object-centric** states from raw RGB inputs. They consist of a list of detected objects with different properties.

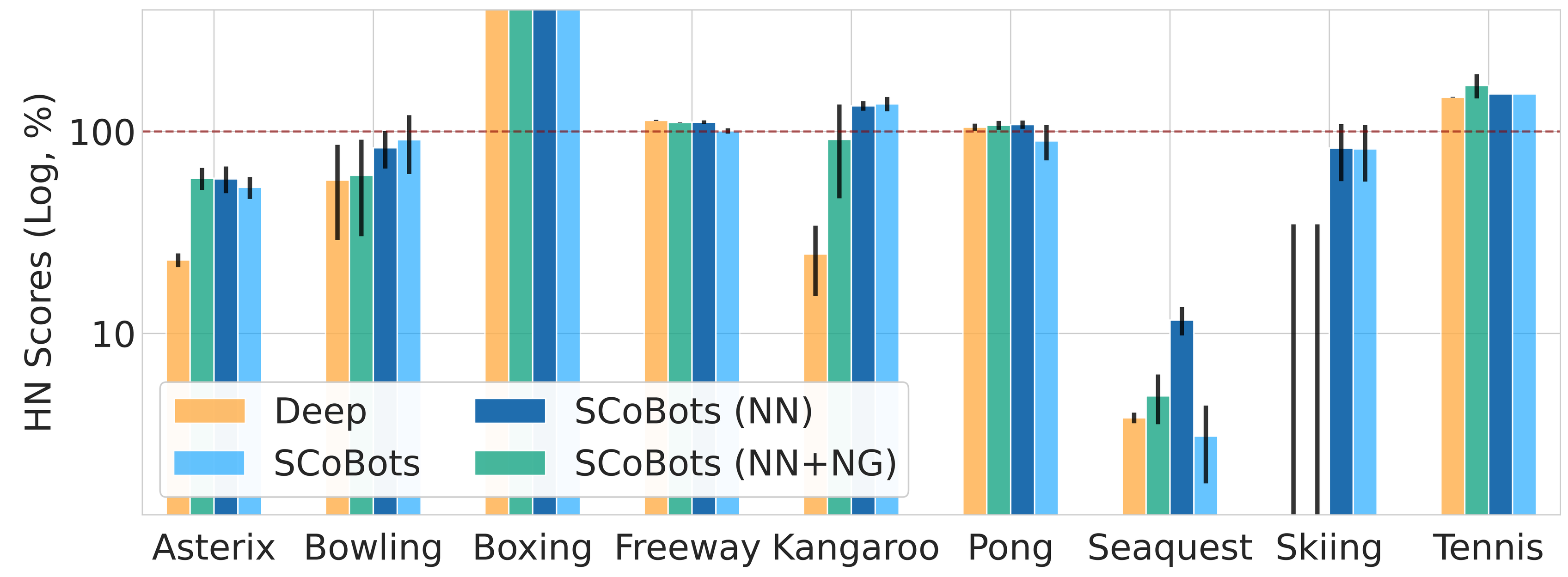
(ii) Relation extractor: Extracts higher level **interpretable relationships** between objects (e.g. speed, distance). These relashions can be provided by a **domain expert** or by an context-provided **LLM**.

(iii) Interpretable Policy Learning: a neural policy is learned, mapping the **relations to actions**, from which an **interpretable decision tree** based policy is extracted

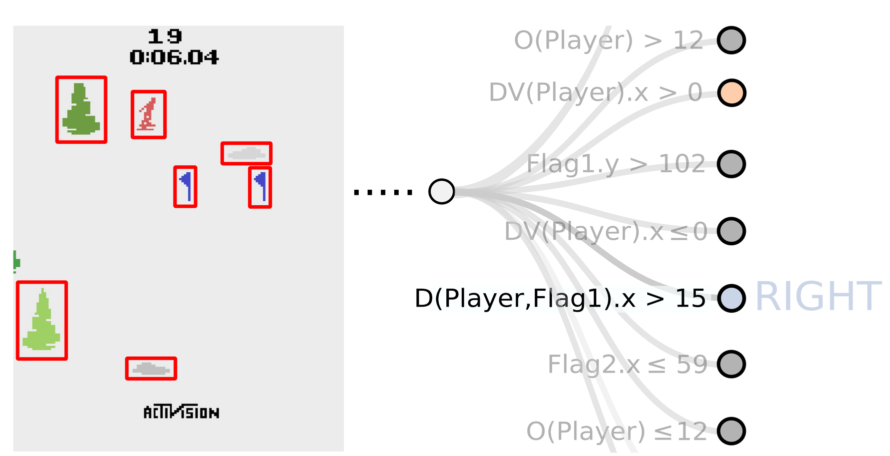
Guiding and correcting SCoBots:

- **prune interpretbale concepts and relations** to prevent the agent from learning shortcuts,
- create **additional reward signals**, using the **interpretable** extracted **concepts**, to help agents with sparse reward or difficult credit assignment.

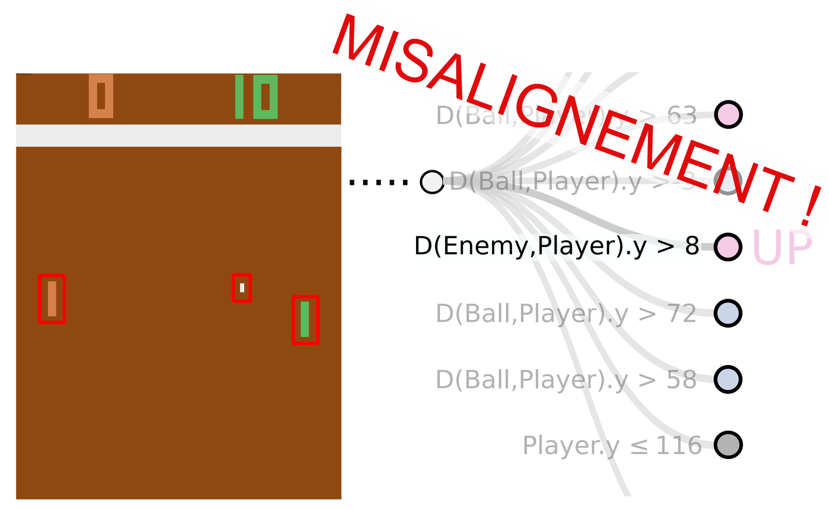
Results: Robust competitive SCoBots



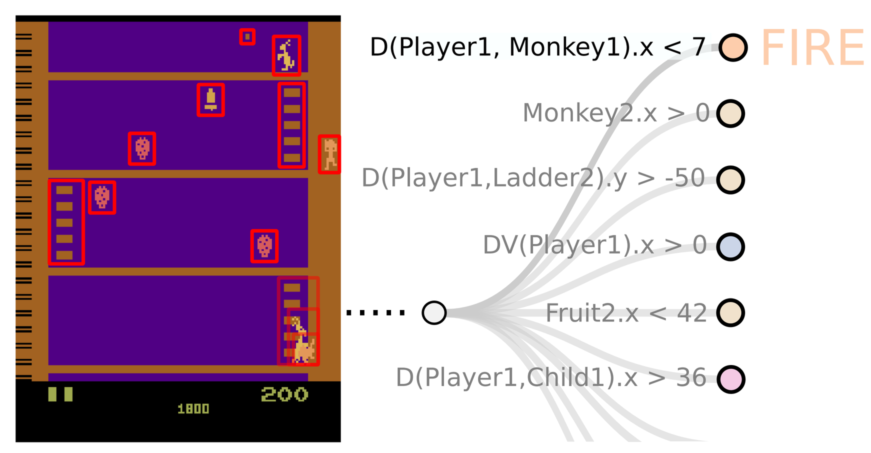
SCoBots can **match or surpass Deep agents performances**, both using **neural networks** or **decision trees**, **with** or without guidance (particularly helping in *Skiing*), shown using human normalized scores on **9 different OCAtari environments** [1].



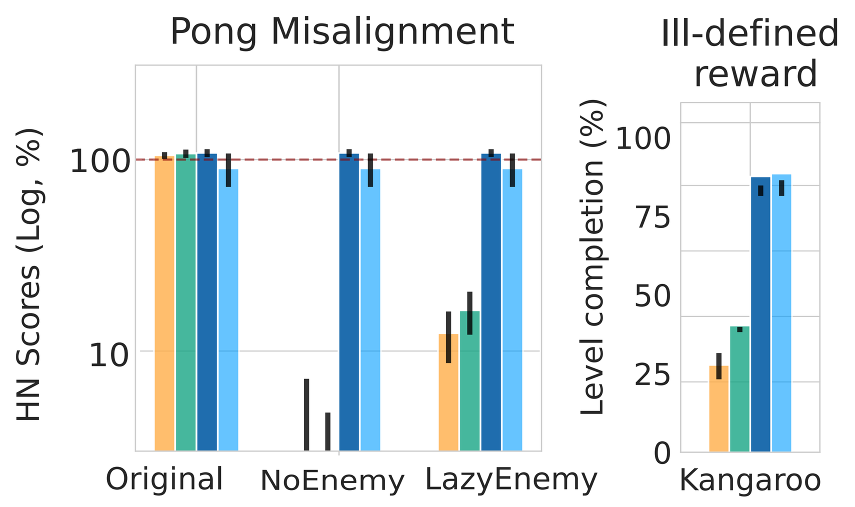
The Player is moving RIGHT as its distance to the Flag (on the x axis) is positive and high.



The Player is moving UP as its distance to the Enemy (on the y axis) is positive.



The Player is punching (FIRE) as its distance to Monkey1 (on the x axis) is positiv and small.



The concept bottlenecks allow for the agents' misalignment correction in Pong and guidance to the intended goal in Kangaroo.

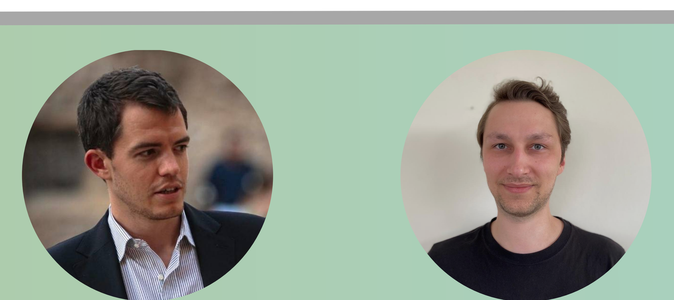
Conclusion

Deep agents are misaligned, as shown in [2] on their tested Atari games.

We introduce **SCoBots, intepretable neurosymbolic RL agents**, that incorporate **successive inspectable concept bottleneck.**

SCoBots allows to **uncover** and **correct misalignments**, as well as for **guidance** to help with **misdefined objectives** and **sparse reward**.

[1] Delfosse, et al. "Ocatari: Object-centric atari 2600 reinforcement learning environments." (2023)
[2] Delfosse, et al. "HackAtari: Atari Learning Environments for Robust and Continual RL." (2024)



Quentin Delfosse
Sebastian Sztwiertnia

[1] AIML Lab TU Darmstadt
[2] ATHENE
[3] hessian.AI
[4] TUDa Centre for Cognitive Science
[5] German Research Center for AI