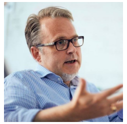
Deep RL agents perform hidden Shortcut Reinforcement Learning. RL agents must rely on human understandable concepts.

Interpretable Concept Bottlenecks to Align Reinforcement Learning Agents

Quentin Delfosse

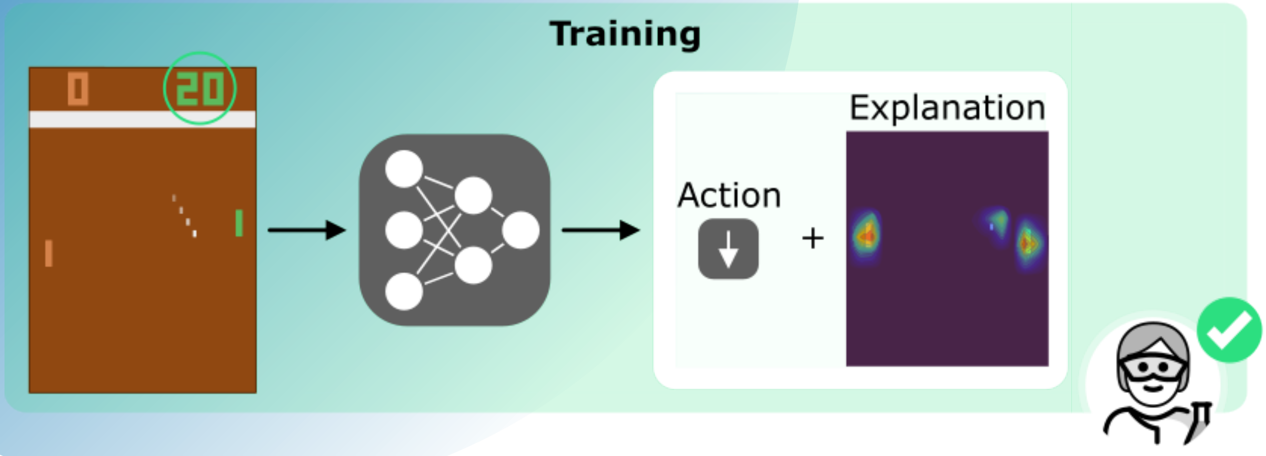
Sebastian Sztwiertnia

Wolfgang Stammer

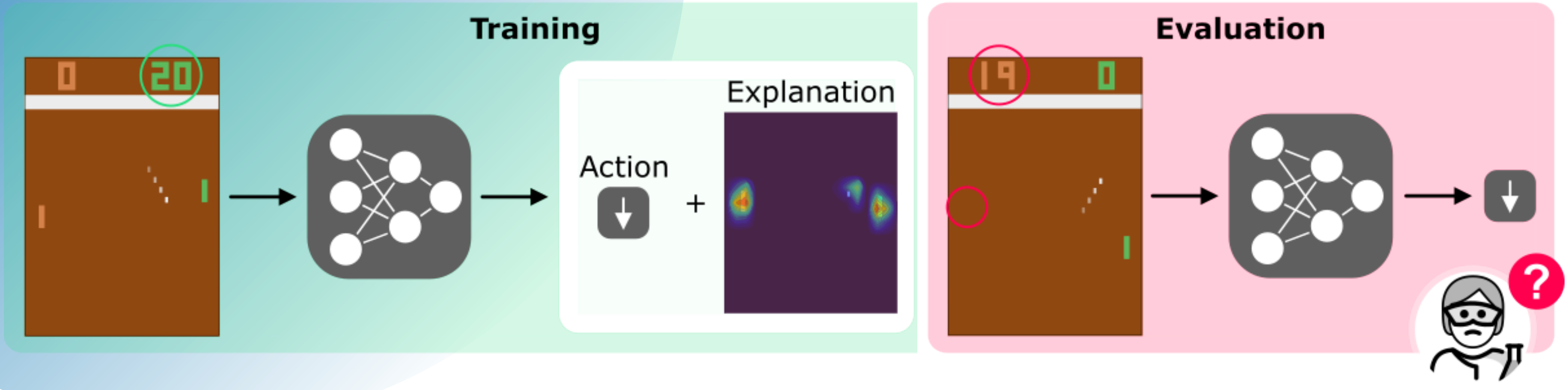


Kristian Kersting

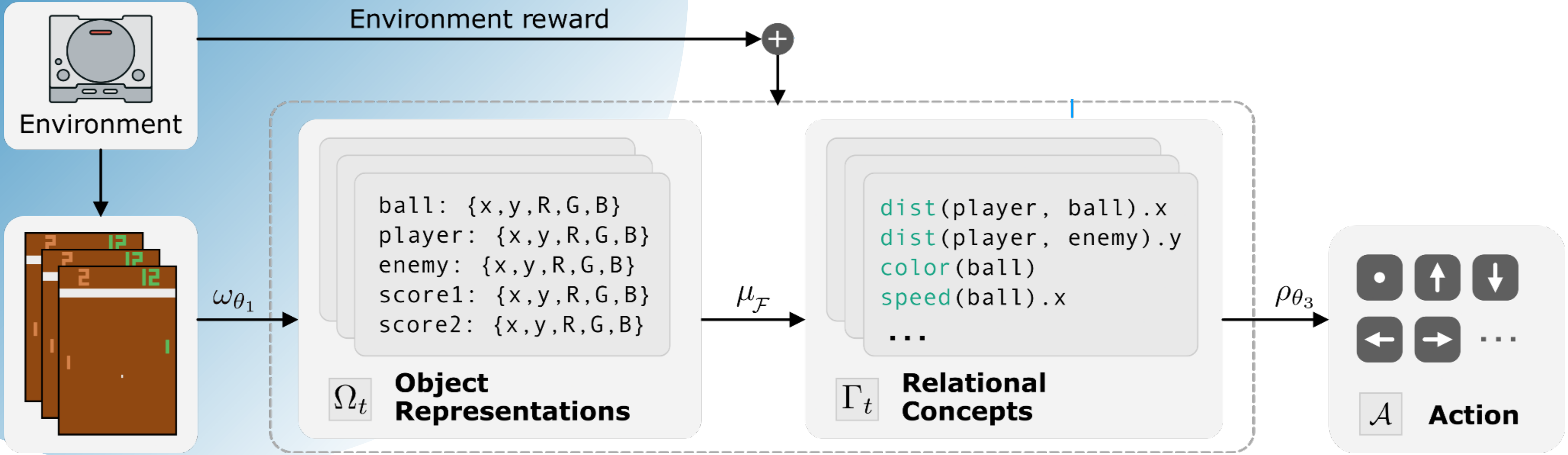
Can we interpret deep agents?



Can we interpret deep agents?

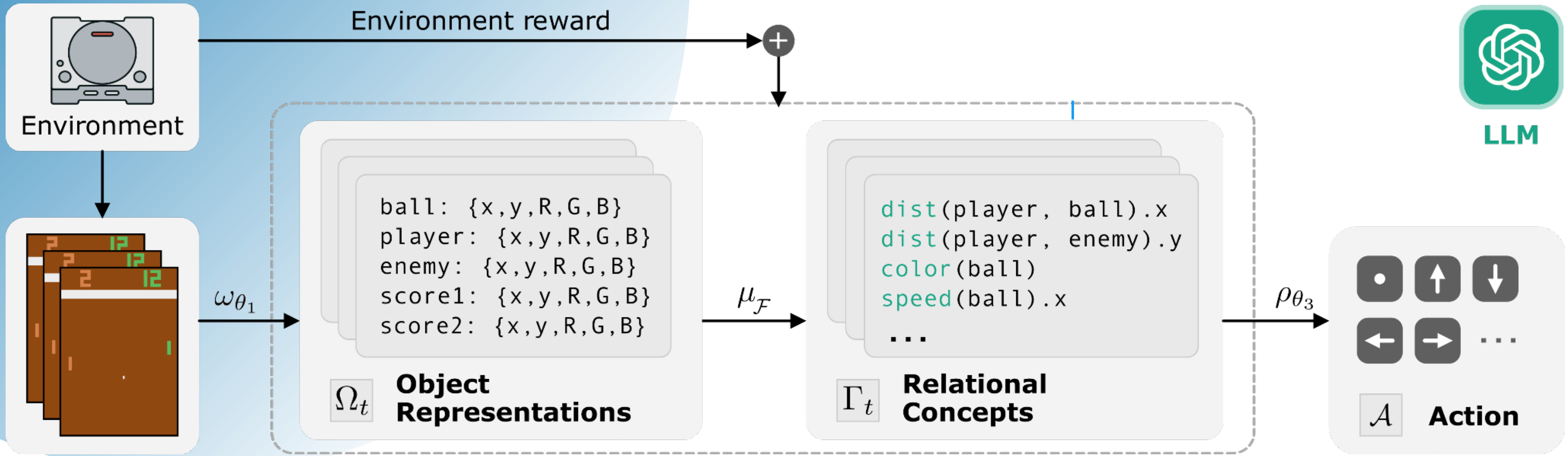


Successive Concept Bottlenecks



SPACE: Lin et al. "SPACE: Unsupervised Object Scene Representation via Spatial Attention and Decomposition." (2020)
SPOC: Delfosse et al. "Boosting object representation learning via motion and object continuity." (2023).

Successive Concept Bottlenecks



Environment reward

Environment

ω_{θ_1}

```
ball: {x,y,R,G,B}
player: {x,y,R,G,B}
enemy: {x,y,R,G,B}
score1: {x,y,R,G,B}
score2: {x,y,R,G,B}
```

Ω_t **Object Representations**

$\mu_{\mathcal{F}}$

```
dist(player, ball).x
dist(player, enemy).y
color(ball)
speed(ball).x
...
```

Γ_t **Relational Concepts**

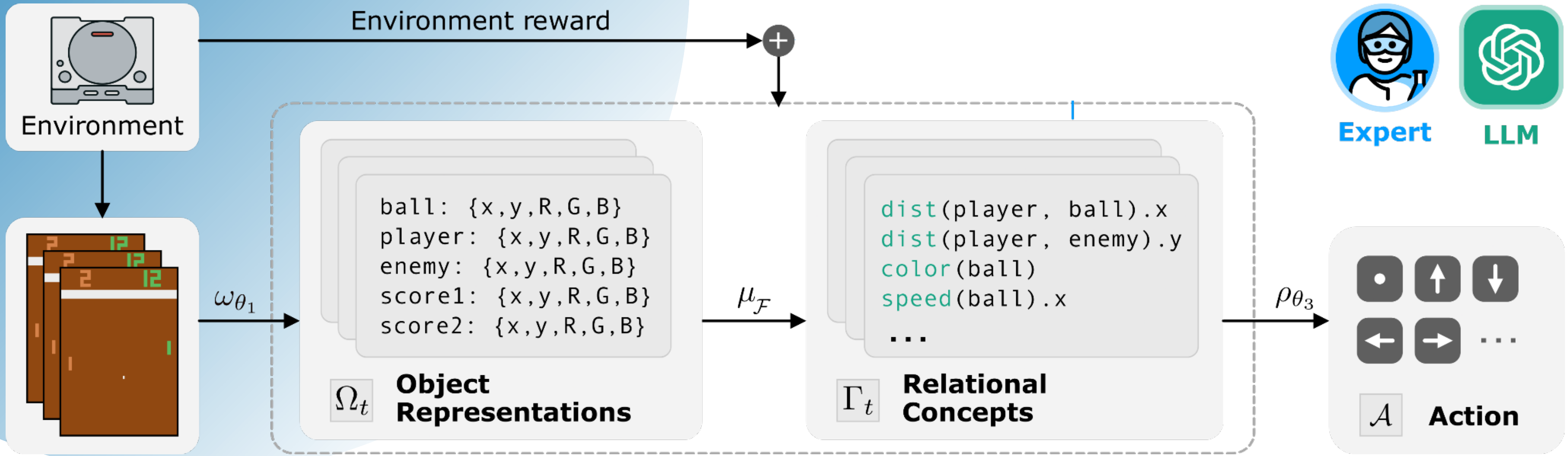
ρ_{θ_3}

$\mathcal{A}$ **Action**

LLM

SPACE: Lin et al. "SPACE: Unsupervised Object Scene Representation via Spatial Attention and Decomposition." (2020)
SPOC: Delfosse et al. "Boosting object representation learning via motion and object continuity." (2023).

Successive Concept Bottlenecks



Environment reward

Environment

ω_{θ_1}

```
ball: {x,y,R,G,B}
player: {x,y,R,G,B}
enemy: {x,y,R,G,B}
score1: {x,y,R,G,B}
score2: {x,y,R,G,B}
```

Ω_t **Object Representations**

$\mu_{\mathcal{F}}$

```
dist(player, ball).x
dist(player, enemy).y
color(ball)
speed(ball).x
...
```

Γ_t **Relational Concepts**

Expert

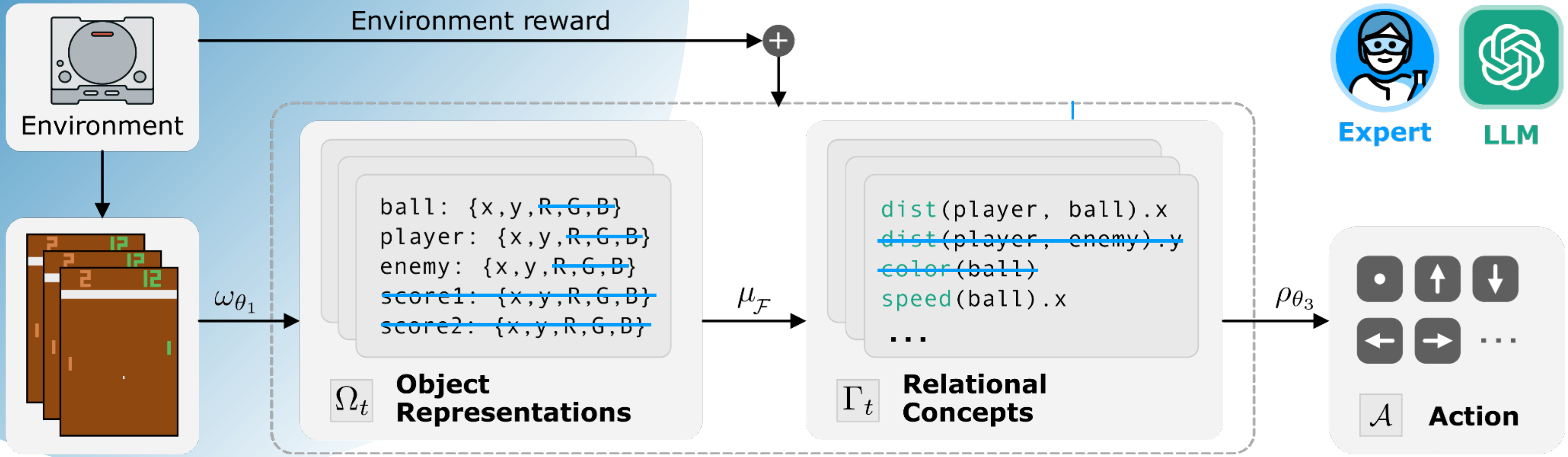
LLM

ρ_{θ_3}

$\mathcal{A}$ **Action**

SPACE: Lin et al. "SPACE: Unsupervised Object Scene Representation via Spatial Attention and Decomposition." (2020)
SPOC: Delfosse et al. "Boosting object representation learning via motion and object continuity." (2023).

Successive Concept Bottlenecks



Environment reward

Environment

ω_{θ_1}

ball: {x,y,~~R,G,B~~}
player: {x,y,~~R,G,B~~}
enemy: {x,y,~~R,G,B~~}
~~score1: {x,y,R,G,B}~~
~~score2: {x,y,R,G,B}~~

Ω_t **Object Representations**

$\mu_{\mathcal{F}}$

dist(player, ball).x
~~dist(player, enemy).y~~
~~color(ball)~~
speed(ball).x
...

Γ_t **Relational Concepts**

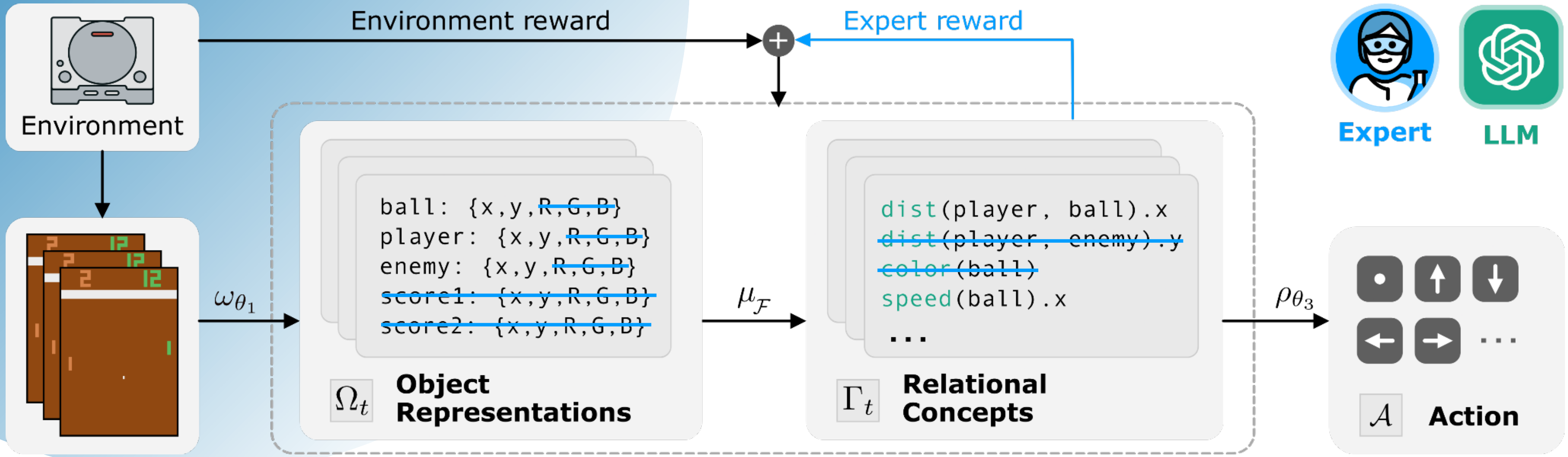
ρ_{θ_3}

$\mathcal{A}$ **Action**

Expert **LLM**

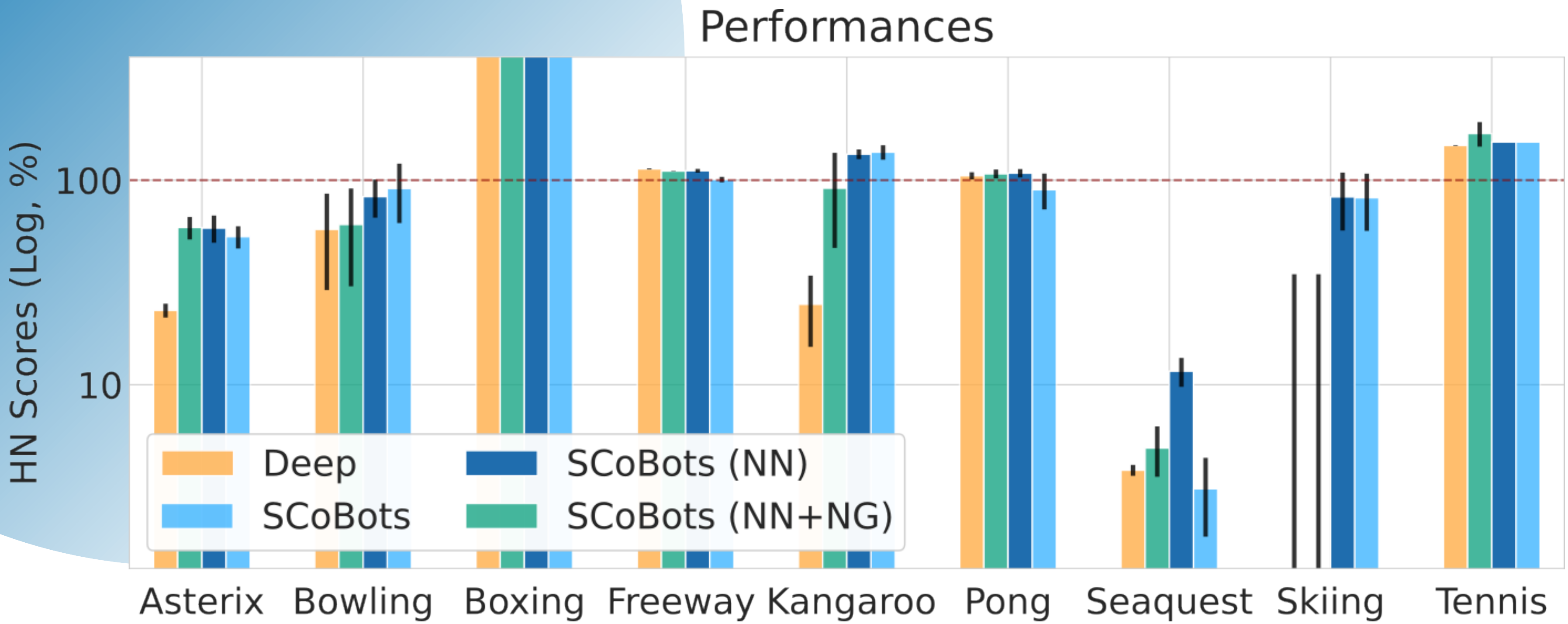
SPACE: Lin et al. "SPACE: Unsupervised Object Scene Representation via Spatial Attention and Decomposition." (2020)
SPOC: Delfosse et al. "Boosting object representation learning via motion and object continuity." (2023).

Successive Concept Bottlenecks

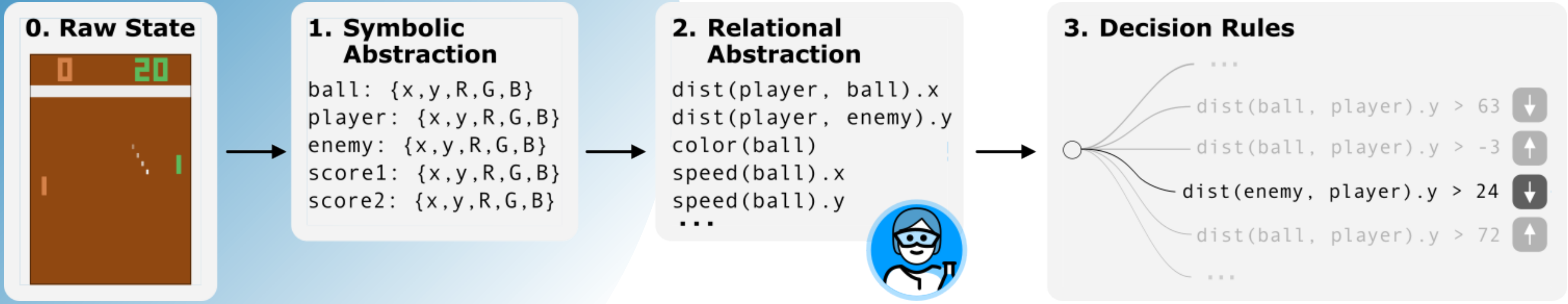


SPACE: Lin et al. "SPACE: Unsupervised Object Scene Representation via Spatial Attention and Decomposition." (2020)
SPOC: Delfosse et al. "Boosting object representation learning via motion and object continuity." (2023).

SCoBots are competitive



SCoBots' can be realigned!



0. Raw State

1. Symbolic Abstraction

ball: {x,y,R,G,B}
player: {x,y,R,G,B}
enemy: {x,y,R,G,B}
score1: {x,y,R,G,B}
score2: {x,y,R,G,B}

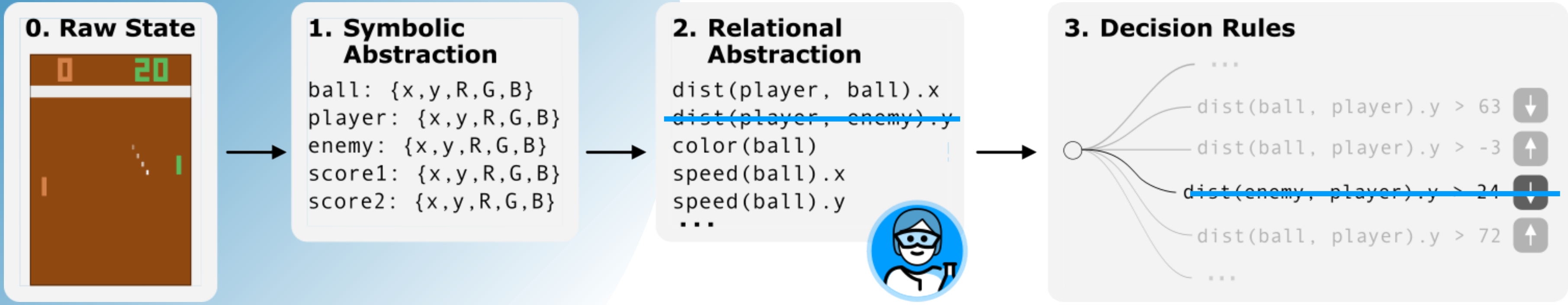
2. Relational Abstraction

dist(player, ball).x
dist(player, enemy).y
color(ball)
speed(ball).x
speed(ball).y
...

3. Decision Rules

...
dist(ball, player).y > 63 ↓
dist(ball, player).y > -3 ↑
dist(enemy, player).y > 24 ↓
dist(ball, player).y > 72 ↑
...

SCoBots' can be realigned!



0. Raw State

1. Symbolic Abstraction

```
ball: {x,y,R,G,B}
player: {x,y,R,G,B}
enemy: {x,y,R,G,B}
score1: {x,y,R,G,B}
score2: {x,y,R,G,B}
```

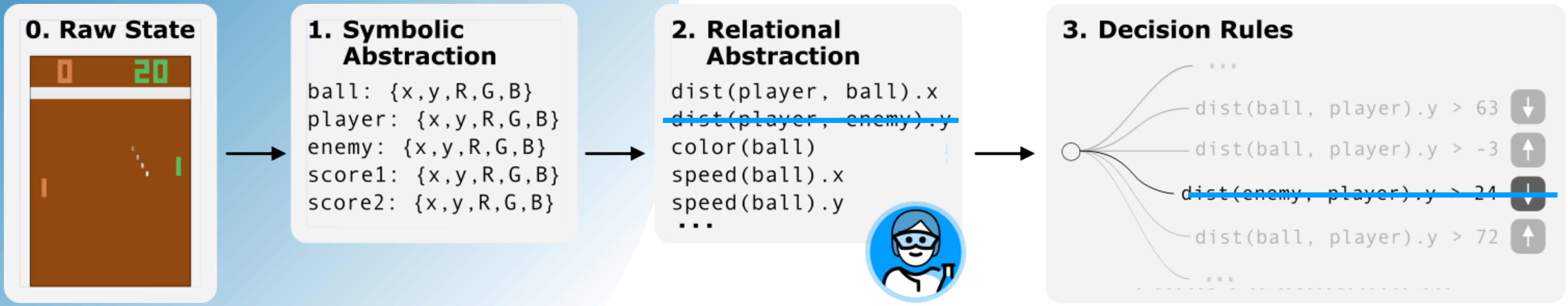
2. Relational Abstraction

dist(player, ball).x
~~dist(player, enemy).y~~
color(ball)
speed(ball).x
speed(ball).y
...

3. Decision Rules

...
dist(ball, player).y > 63
dist(ball, player).y > -3
~~dist(enemy, player).y > 24~~
dist(ball, player).y > 72
...

SCoBots' can be realigned!



0. Raw State

1. Symbolic Abstraction

```
ball: {x,y,R,G,B}
player: {x,y,R,G,B}
enemy: {x,y,R,G,B}
score1: {x,y,R,G,B}
score2: {x,y,R,G,B}
```

2. Relational Abstraction

dist(player, ball).x

~~dist(player, enemy).y~~

color(ball)

speed(ball).x

speed(ball).y

...

3. Decision Rules

...

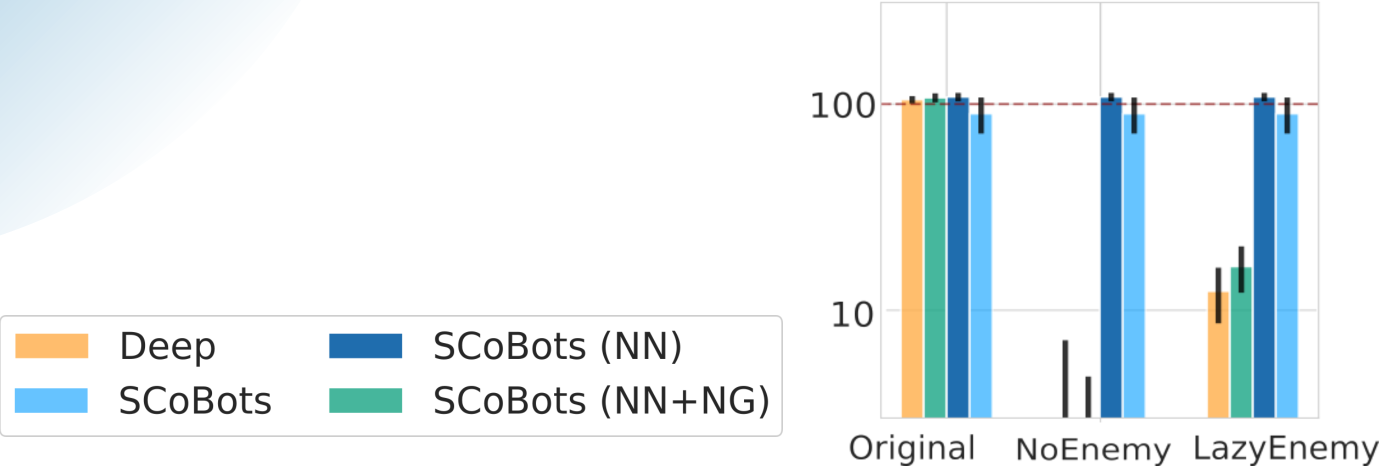
dist(ball, player).y > 63

dist(ball, player).y > -3

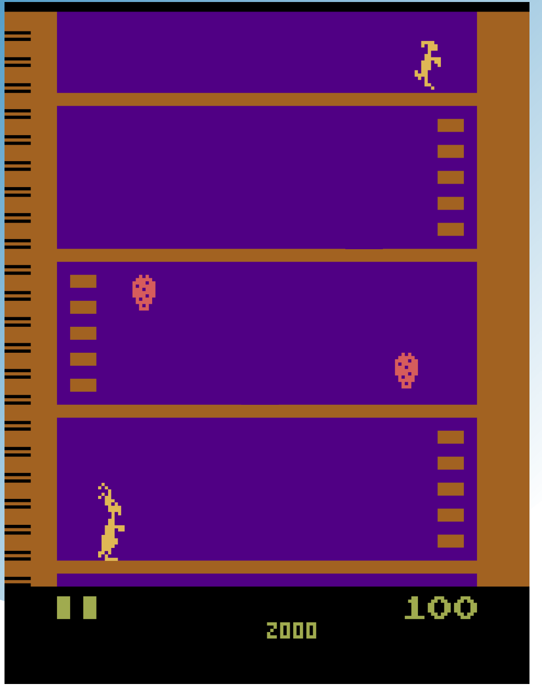
~~dist(enemy, player).y > 24~~

dist(ball, player).y > 72

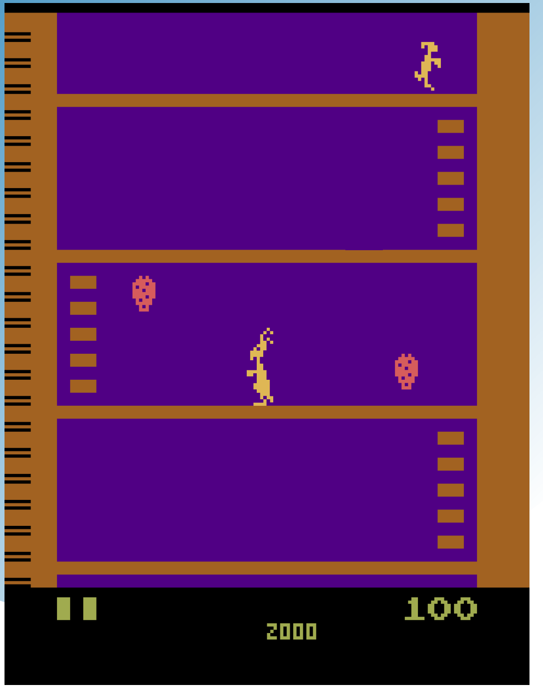
...



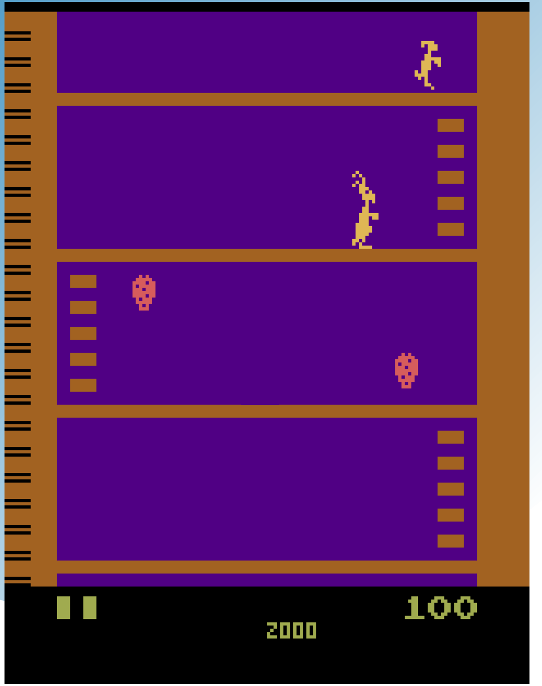
SCoBots can be redirected!



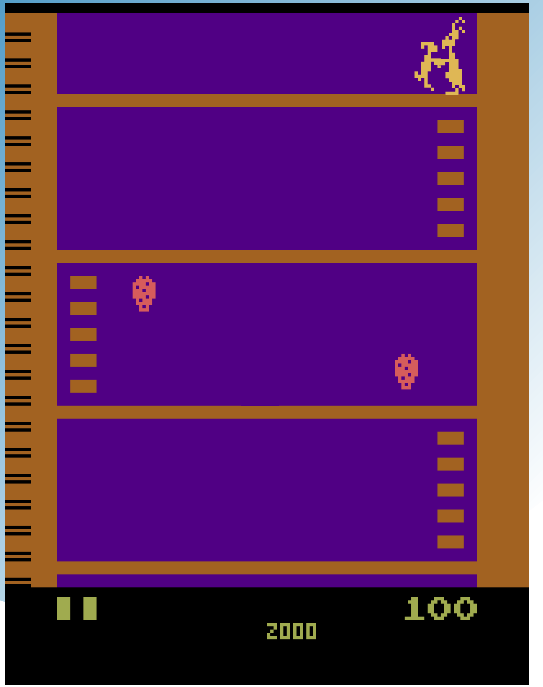
SCoBots can be redirected!



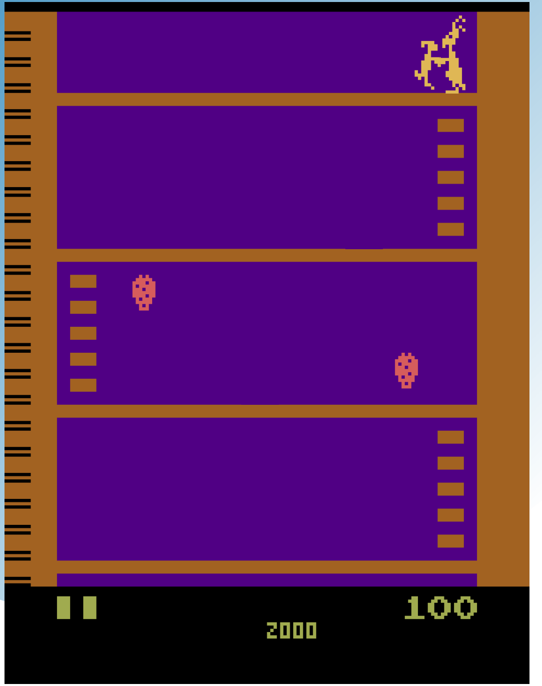
SCoBots can be redirected!

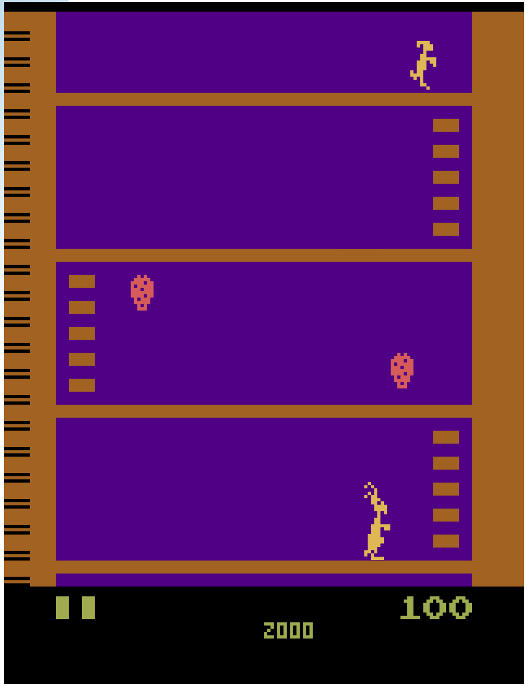


SCoBots can be redirected!

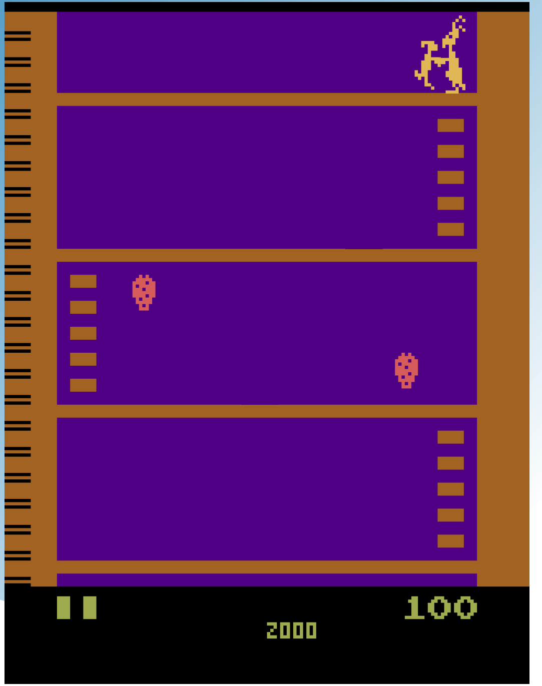


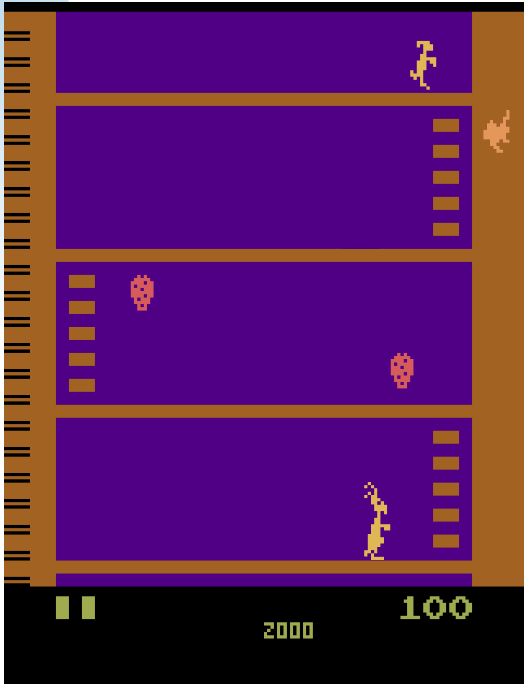
SCoBots can be redirected!



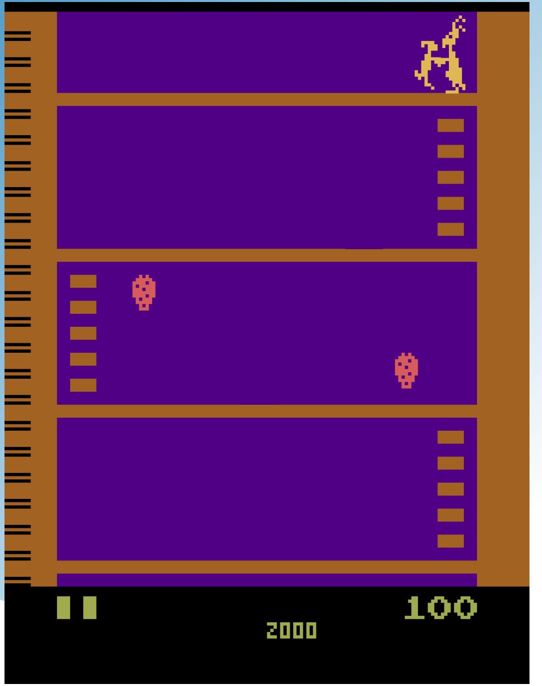


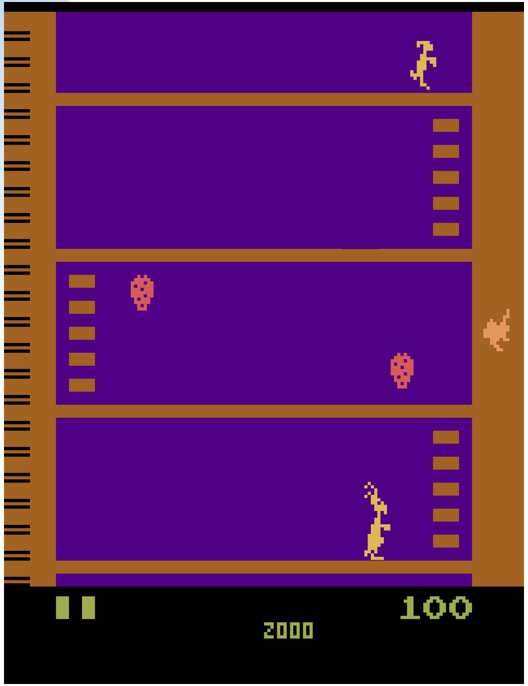
SCoBots can be redirected!



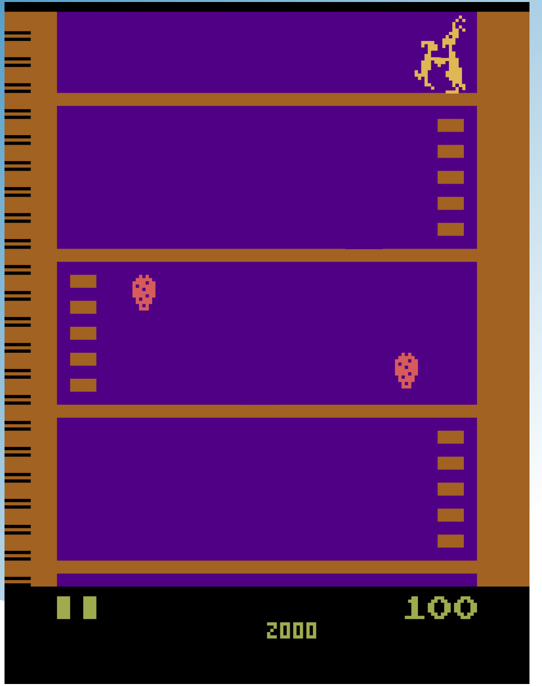


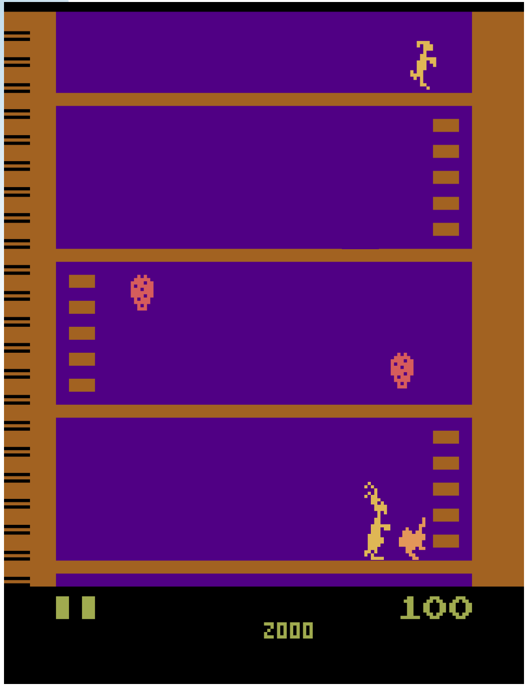
SCoBots can be redirected!



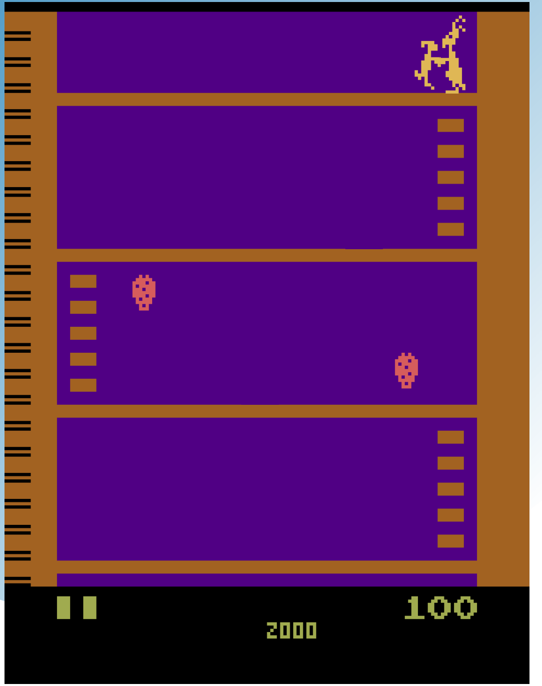


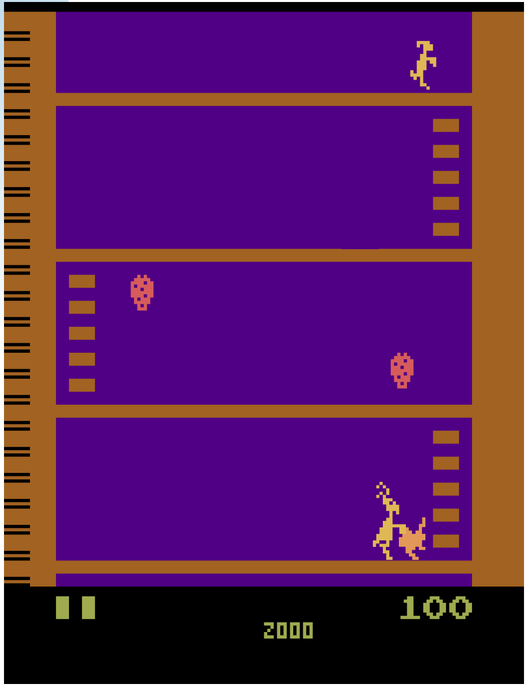
SCoBots can be redirected!



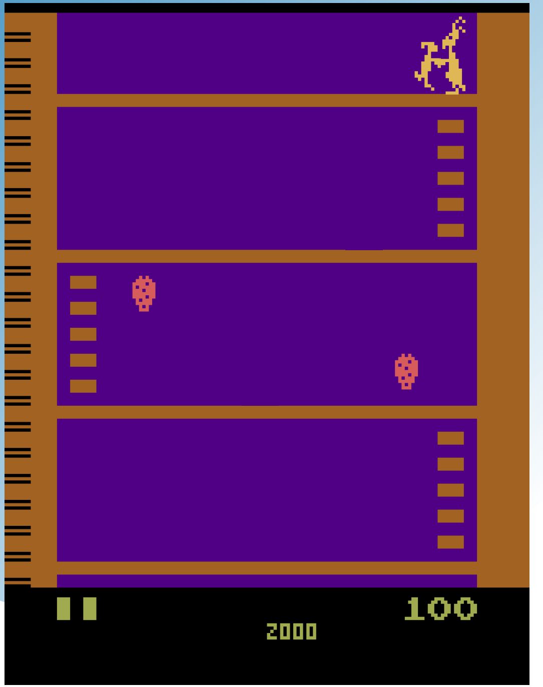


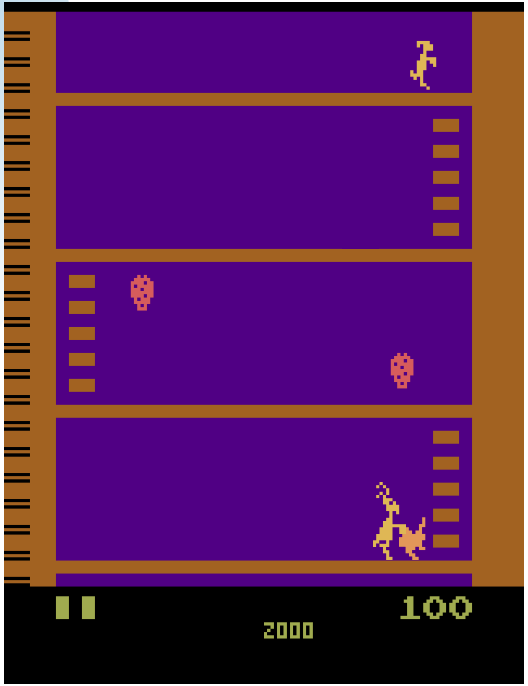
SCoBots can be redirected!

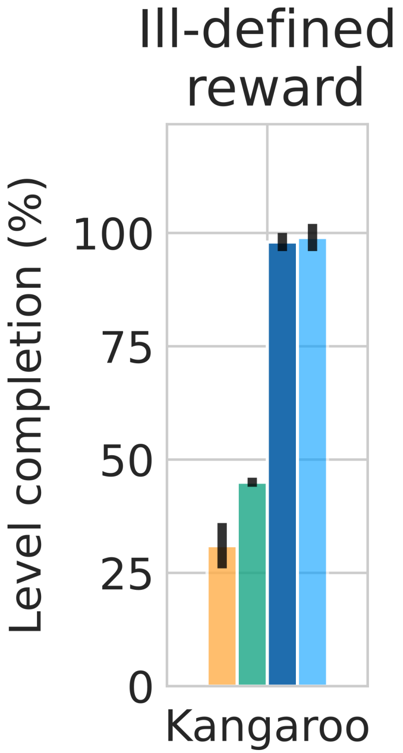




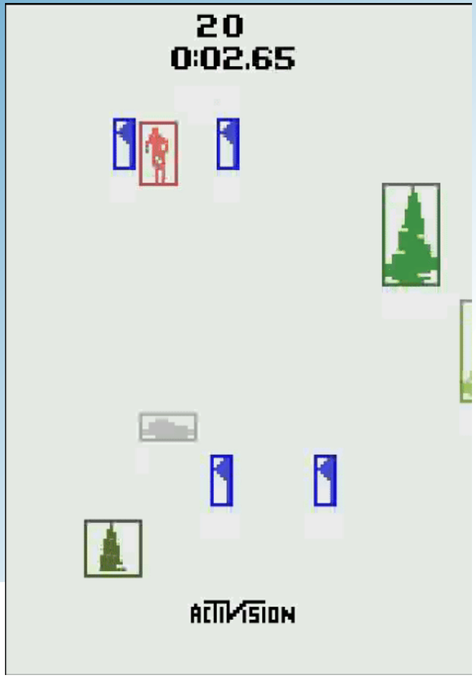
SCoBots can be redirected!

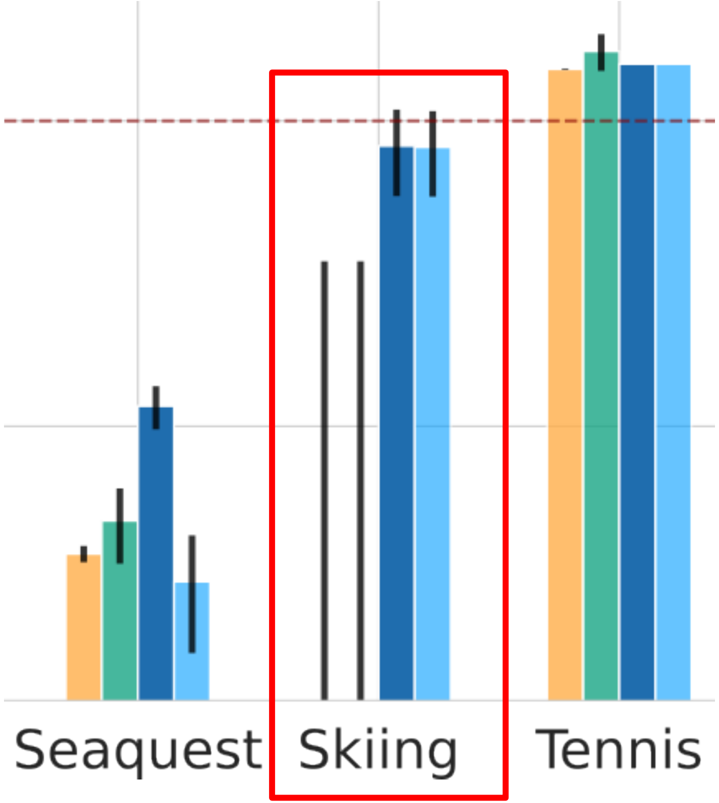




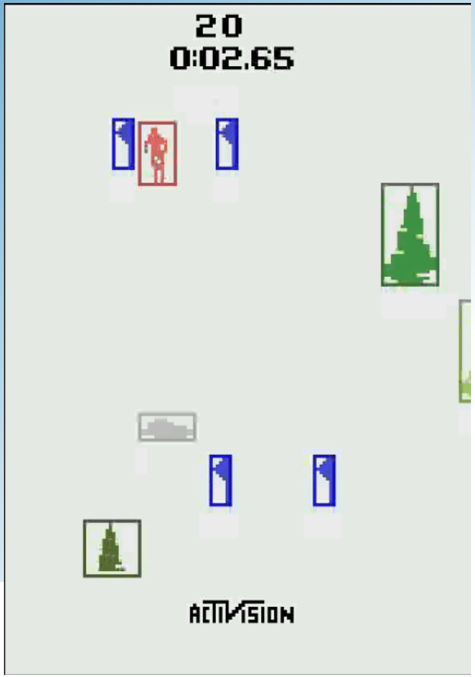


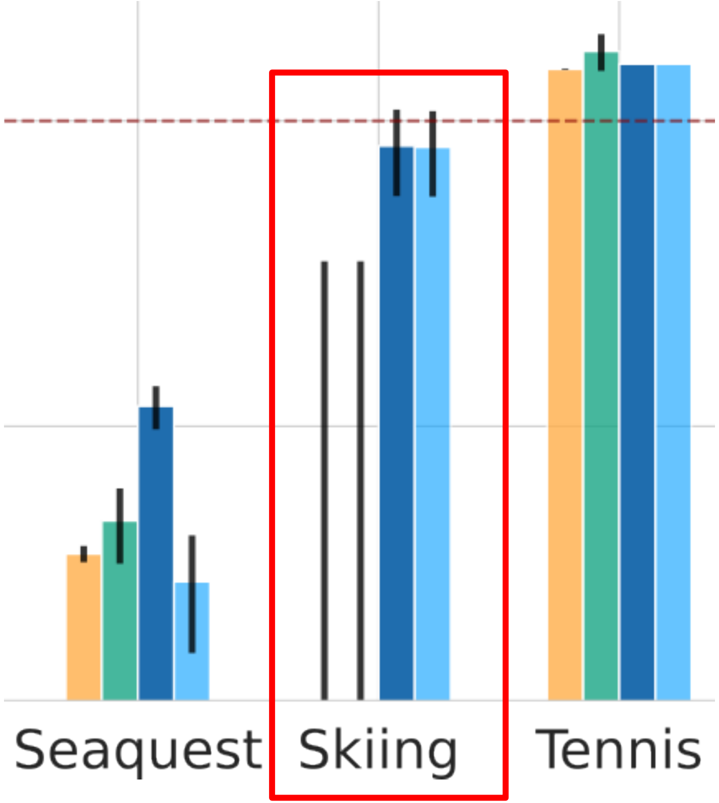
SCoBots can be redirected





SCoBots can be redirected





Deep RL agents perform hidden Shortcut Reinforcement Learning. RL agents must rely on human understandable concepts.

Open source: github.com/k4ntz/SCoBots

Reinforcement
Learning
Conference